Practical Posterior Error Bounds from Variational Objectives

Jonathan H. Huggins [1] **Mikolaj Kasprzak** [2] **Trevor Campbell** [3] **Tamara Broderick** [4]

Abstract

Variational inference has become an increasingly attractive fast alternative to Markov chain Monte Carlo methods for approximate Bayesian inference. However, a major obstacle to the widespread use of variational methods is the lack of post-hoc accuracy measures that are both theoretically justified and computationally efficient. In this paper, we provide rigorous bounds on the error of posterior mean and uncertainty estimates that arise from full-distribution approximations, as in variational inference. Our bounds are widely applicable as they require only that the approximating and exact posteriors have polynomial moments. Our bounds are computationally efficient for variational inference in that they require only standard values from variational objectives, straightforward analytic calculations, and simple Monte Carlo estimates. We show that our analysis naturally leads to a new and improved workflow for variational inference. Finally, we demonstrate the utility of our proposed workflow and error bounds on a real-data example with a widely used multilevel hierarchical model.

1. Introduction

Exact Bayesian statistical inference is known for providing point estimates with desirable decision-theoretic properties as well as coherent uncertainties. Using Bayesian methods in practice, though, typically requires approximating these quantities. It is crucial, then, to quantify the error introduced by any approximation. There are two, essentially complementary, options: (1) rigorous a priori characterization of accuracy for finite data and (2) tools for evaluating approximation accuracy a posteriori. First, consider option #1. Markov chain Monte Carlo (MCMC) methods are the gold standard for sound approximate Bayesian inference in part due to their flexibility and strong a priori theoretical guarantees on quality for finite data. However, these guarantees are typically asymptotic in running time, and computational concerns have motivated a spate of alternative Bayesian approximations. Within the machine learning community, variational approaches (Blei et al., 2017; Wainwright et al., 2008) such as black-box and automatic differentiation variational inference (Kucukelbir et al., 2015; Ranganath et al., 2014) are perhaps the most widely used. While these methods have empirically demonstrated computational gains on problems of interest, they do not come equipped with guarantees on the approximation accuracy of point estimates and uncertainties. There has been some limited but ongoing work in developing relevant a priori guarantees for common variational approaches (Alquier and Ridgway, 2017; Alquier et al., 2016; Chérief-Abdellatif and Alquier, 2018; Pati et al., 2018; Wang and Blei, 2018). There has also been work in developing new (boosting) variational algorithms that come equipped with a priori guarantees on the convergence of the approximation distribution to arbitrary accuracy (Campbell and Li, 2019; Guo et al., 2016; Locatello et al., 2018a;b; Miller et al., 2017; Wang, 2016).

The examples above typically either have no guarantees or purely asymptotic guarantees – or require non-convex optimization. Thus, in every case, reliable evaluation tools (option #2) would provide an important bulwark for data analysis. In any particular data analysis, such tools could determine if the approximate point estimates and uncertainties are to be trusted. Gorham and Mackey (2015; 2017); Gorham et al. (2019); Yao et al. (2018) have pioneered initial work in developing evaluation tools applicable to variational inference. However, current methods are either heuristic or computationally inefficient. They may also have intractable constants or impractically strong assumptions on tail behavior.

In this paper, we provide the first rigorous *and* computationally efficient error bounds on the quality of posterior point and uncertainty estimates for variational approximations. We highlight three practical aspects of our bounds here: (A) computational efficiency, (B) weak tail restrictions, and (C) relevant targets. For A, we use only stan-

[1]Department of Biostatistics, Harvard T. H. Chan School of Public Health [2]Mathematics Research Unit, University of Luxembourg [3]Department of Statistics, University of British Columbia [4]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology. Correspondence to: Jonathan H. Huggins <jhuggins@mit.edu>.

dard values computed in the course of variational inference, straightforward analytic calculations, and simple Monte Carlo (not MCMC) estimates. For B, we require only that the approximating and exact posteriors have polynomial moments – though we show even tighter bounds when exponential moments exist. For C, note that practitioners typically report posterior means for point estimates – and they report posterior variance, standard deviation, or mean absolute deviation for uncertainties (Gelman et al., 2013; Robert, 1994). So we directly bound the error in these quantities. We demonstrate the importance of bounding error in these output quantities directly, rather than bounding divergences between distributions, with illustrative counterexamples; namely, we show that common variational objectives such as the Kullback–Leibler (KL) divergence and α-divergences can be very small at the same time that mean and variance estimates are *arbitrarily* wrong.

To obtain our bounds, we make three main technical contributions, which may all be of independent interest beyond Bayesian methods. First, we show how to bound mean and uncertainty differences in terms of Wasserstein distance. Second, we develop novel bounds on the Wasserstein distance in terms of α-divergences – including the KL divergence – and moment bounds on the variational approximation. The moment conditions allow us to relate (scale-free) α-divergences to (scale-sensitive) Wasserstein distances. Finally, we derive efficiently computable bounds on α-divergences in terms of the objectives already widely used for variational inference – in particular, the evidence lower bound (ELBO) and χ upper bound (CUBO) (Dieng et al., 2017). By combining all three contributions, we obtain efficiently computed bounds on means and uncertainties in terms of the ELBO, CUBO, and certain polynomial or exponential moments of the variational approximation. Our methods give rise to a new and improved workflow for variational inference. We illustrate the usefulness of our bounds as well as the practicality of our new workflow on a real-data example with a widely used multilevel hierarchical model.[1]

2. Preliminaries

Bayesian inference. Let $\theta \in \mathbb{R}^d$ denote a parameter vector of interest, and let z denote observed data. A Bayesian model consists of a prior measure $\pi_0(\mathrm{d}\theta)$ and a likelihood $\ell(z;\theta)$. Together, the prior and likelihood define a joint distribution over the data and parameters. The Bayesian posterior distribution π is the conditional in θ with fixed data z.[2] To write this conditional, we define the proportional posterior measure $\pi'(\mathrm{d}\theta) := \ell(z;\theta)\pi_0(\mathrm{d}\theta)$, and the marginal likelihood, or evidence, $M := \int \mathrm{d}\pi'$. Then the posterior is $\pi := \pi'/M$.

Typically, practitioners are concerned with the quality of a number of *summaries*—point estimates and uncertainties—of the posterior approximation compared with those of the exact posterior. For a given approximate distribution $\hat{\pi}$ on $\mathbb{R}^d$, such summaries include the mean $m_{\hat{\pi}}$, covariance $\Sigma_{\hat{\pi}}$, ith component marginal standard deviation $\sigma_{\hat{\pi},i}$, and mean absolute deviation $\mathrm{MAD}_{\hat{\pi},i}$: for $\vartheta \sim \hat{\pi}$,

$$
\begin{aligned}
m_{\hat{\pi}} &:= \mathbb{E}(\vartheta), \quad \mathrm{MAD}_{\hat{\pi},i} := \mathbb{E}(|\vartheta_i - m_{\hat{\pi},i}|), \\
\sigma_{\hat{\pi},i} &:= \Sigma_{\hat{\pi},ii}^{1/2}, \qquad \Sigma_{\hat{\pi}} := \mathbb{E}\{(\vartheta - m_{\hat{\pi}})(\vartheta - m_{\hat{\pi}})^\top\}.
\end{aligned}
$$

Variational inference. In most applications of interest, it is infeasible to efficiently compute these summaries with respect to the posterior distribution in closed form or via simple Monte Carlo. Therefore, one must use an approximate inference method, which produces an approximation $\hat{\pi}$ to the posterior π. One approach, *variational inference*, is widely used in machine learning. Variational inference aims to minimize some measure of discrepancy $\mathcal{D}_\pi(\cdot)$ over a tractable family $\mathcal{Q}$ of potential approximation distributions (Blei et al., 2017; Wainwright et al., 2008):

$$
\hat{\pi} = \operatorname*{arg\,min}_{\xi \in \mathcal{Q}} \mathcal{D}_\pi(\xi).
$$

The variational family $\mathcal{Q}$ is chosen to be tractable in the sense that, for any $\xi \in \mathcal{Q}$, we are able to efficiently calculate relevant summaries either analytically or using independent and identically distributed samples from ξ.

KL divergence. The classical choice for the discrepancy in variational inference is the *Kullback–Leibler (KL) divergence* (or *relative entropy*) (Bishop, 2006):

$$
\mathrm{KL}(\xi \mid \pi) := \int \log\left(\frac{\mathrm{d}\xi}{\mathrm{d}\pi}\right)\mathrm{d}\xi.
$$

Note that the KL divergence is asymmetric in its arguments. The direction $\mathcal{D}_\pi(\xi) = \mathrm{KL}(\xi \mid \pi)$ is most typical in variational inference, largely out of convenience; the unknown marginal likelihood M appears in an additive constant that does not influence the optimization, and computing gradients requires estimating expectations only with respect to $\xi \in \mathcal{Q}$, which is chosen to be tractable. Minimizing $\mathrm{KL}(\xi \mid \pi)$ is equivalent to maximizing the *evidence lower bound*, or $\mathsf{ELBO}(\xi)$ (Bishop, 2006):

$$
\mathsf{ELBO}(\xi) := \int \log\left(\frac{\mathrm{d}\pi'}{\mathrm{d}\xi}\right)\mathrm{d}\xi
$$

[1] A `python` package for computing the bounds we develop in this paper is available at `https://github.com/jhuggins/viabel`.

[2] Since the data z are always fixed throughout this work, we have suppressed the dependence on z in the notation.

Rényi's α-divergence. Another choice of discrepancy for variational inference (Bui et al., 2017; Dieng et al., 2017; Hernández-Lobato et al., 2016; Li and Turner, 2016) is Rényi's α-divergence, which for $\alpha \in (0,1) \cup (1,\infty)$ is defined as

$$\mathrm{D}_\alpha(\pi \mid \xi) := \frac{1}{\alpha-1}\log\int\left(\frac{\mathrm{d}\pi}{\mathrm{d}\xi}\right)^{\alpha-1}\mathrm{d}\pi.$$

The α-divergence is typically used in variational inference with $\mathcal{D}_\pi(\xi) = \mathrm{D}_\alpha(\pi \mid \xi)$ for $\alpha > 1$; again, the unknown marginal likelihood M does not influence the optimization, and estimating gradients is tractable. Variational inference with the α-divergence is equivalent to minimizing a quantity known as the χ *upper bound*, or $\mathsf{CUBO}_\alpha(\xi)$ (Dieng et al., 2017), which is equal to $(1-\alpha^{-1})\mathrm{D}_\alpha(\pi \mid \xi) - \log M$. The ELBO and CUBO are so-named since they respectively provide a lower and upper bound for $\log M$; see Appendix D.2. The α-divergence generalizes the KL divergence since $\mathrm{D}_\alpha(\pi \mid \xi) := \lim_{\alpha\to 1}\mathrm{D}_\alpha(\pi \mid \xi) = \mathrm{KL}(\pi \mid \xi)$ (Cichocki and Amari, 2010). Note, however, that here the KL divergence has the order of its arguments switched when compared to how it is used for variational inference.

Wasserstein distance. The Wasserstein distance is a measure of discrepancy that, unlike the previous two divergences, is influenced by a metric on the underlying space. It is widely used in the analysis of Markov chain Monte Carlo algorithms and large-scale data asymptotics (e.g., Durmus and Moulines, 2019; Durmus et al., 2019; Eberle and Majka, 2019; Joulin and Ollivier, 2010; Madras and Sezer, 2010; Rudolf and Schweizer, 2018; Vollmer et al., 2016). The p-Wasserstein distance between ξ and π is given by

$$\mathcal{W}_p(\xi,\pi) := \inf_{\gamma\in\Gamma(\xi,\pi)}\left\{\int\|\theta-\theta'\|_2^p\gamma(\mathrm{d}\theta,\mathrm{d}\theta')\right\}^{1/p},$$

where $\Gamma(\xi,\pi)$ is the set of *couplings* between ξ and π, i.e., Borel measures γ on $\mathbb{R}^d \times \mathbb{R}^d$ such that $\xi = \gamma(\cdot,\mathbb{R}^d)$ and $\pi = \gamma(\mathbb{R}^d,\cdot)$ (Villani, 2009, Defs. 6.1, 1.1). The Wasserstein distance is difficult to use as a variational objective due to the (generally intractable) infimum over couplings, although there is recent work in this direction (Claici et al., 2018; Cuturi and Doucet, 2014; Srivastava et al., 2018).

3. Error bounds via posterior discrepancies

Given the concern with posterior summaries, a meaningful measure of posterior approximation quality should control the error in each of these summaries, i.e., $\|m_{\hat\pi} - m_\pi\|_2$, $|\,\mathrm{MAD}_{\hat\pi,i} - \mathrm{MAD}_{\pi,i}\,|$, $\|\Sigma_{\hat\pi} - \Sigma_\pi\|_2$, and $|\sigma_{\hat\pi,i} - \sigma_{\pi,i}|$. To be practical, this measure should also be computationally efficient. In this section, we focus only on the challenge of finding a discrepancy that controls the error of these summaries. In particular, we (1) provide counterexamples to show that $\mathrm{KL}(\hat\pi \mid \pi)$ and $\mathrm{D}_\alpha(\pi \mid \hat\pi)$ by themselves cannot be relied upon to control these errors (see Appendix A for full details), and (2) prove that the Wasserstein distance does provide the desired control. We address computational efficiency in Section 4.

KL divergence. Unfortunately, as we show in the following examples, even when $\mathrm{KL}(\hat\pi \mid \pi)$ is small, posterior summary approximations provided by $\hat\pi$ can be arbitrarily poor. The simplicity of our examples suggests that these results are not due merely to exploiting degeneracies. First we note that the exact posterior standard deviation σ_π is a natural scale for the posterior mean error since changing the posterior mean by σ_π or more could fundamentally change practical decisions made based on the posterior. Our first example shows that even when $\mathrm{KL}(\hat\pi \mid \pi)$ is small, the mean error can be arbitrarily large, whether measured relative to σ_π or $\sigma_{\hat\pi}$.

Example 3.1 (Arbitrarily poor mean approximation)**.** For any $t > 0$, there exist (A) one-dimensional, unimodal distributions $\hat\pi$ and π such that $\mathrm{KL}(\hat\pi \mid \pi) < 0.9$ and $(m_{\hat\pi} - m_\pi)^2 > t\sigma_\pi^2$, and (B) one-dimensional, unimodal distributions $\hat\pi$ and π such that $\mathrm{KL}(\hat\pi \mid \pi) < 0.3$ and $(m_{\hat\pi} - m_\pi)^2 > t\sigma_{\hat\pi}^2$. ☐

In more detail, let $\mathsf{Weibull}(k,1)$ denote the Weibull distribution with shape $k > 0$ and scale 1. For (A), we let $\pi = \mathsf{Weibull}(k,1)$, $\hat\pi = \mathsf{Weibull}(k/2,1)$, and $k \searrow 0$. We exchange the two distributions for (B).

Our second example shows that $\mathrm{KL}(\hat\pi \mid \pi)$ can remain small even when the variance difference is arbitrarily large.

Example 3.2 (Arbitrarily poor variance approximation)**.** For any $t \in (1,\infty]$, there exist one-dimensional, mean-zero, unimodal distributions $\hat\pi$ and π such that $\mathrm{KL}(\hat\pi \mid \pi) < 0.23$ but $\sigma_\pi^2 \geq t\sigma_{\hat\pi}^2$. ☐

Here we let $\pi = \mathcal{T}_h$ (standard t-distribution with h degrees of freedom), $\hat\pi = \mathcal{N}(0,1)$ (standard Gaussian), and $h \searrow 2$.

Rényi's α-divergence. We similarly demonstrate that small $\mathrm{D}_\alpha(\pi \mid \hat\pi)$ does not imply accurate mean or variance estimates. We focus on the canonical case $\alpha = 2$, which will also play a key role in our analyses below.

Example 3.3 (Arbitrarily poor mean and variance approximation)**.** For any $t > 0$, there exist two one-dimensional, unimodal distributions $\hat\pi$ and π with $\mathrm{D}_2(\pi \mid \hat\pi) < 0.4$ such that $\sigma_{\hat\pi}^2 \geq t\sigma_\pi^2$ and $(m_{\hat\pi} - m_\pi)^2 \geq t\sigma_\pi^2$. ☐

We again take $\pi = \mathsf{Weibull}(k,1)$, $\hat\pi = \mathsf{Weibull}(k/2,1)$, and $k \searrow 0$.

Wasserstein distance. In contrast to both the KL and α-divergences, the Wasserstein distance accounts for the metric on the underlying space. Intuitively, the Wasserstein distance is large when the mass of two distributions is "far apart." Thus, it is a natural choice of discrepancy for bounding the error in the approximate posterior mean and uncertainty, since these quantities also depend on the underlying metric. Our next result confirms that the Wasserstein distance controls the error in these quantities.

Theorem 3.1. *If $\mathcal{W}_1(\hat{\pi}, \pi) \le \varepsilon$, then*

$$\|m_{\hat{\pi}} - m_\pi\|_2 \le \varepsilon \ \ \textit{and} \ \ \max_i |\,\mathrm{MAD}_{\hat{\pi},i} - \mathrm{MAD}_{\pi,i}\,| \le 2\varepsilon.$$

If $\mathcal{W}_2(\hat{\pi}, \pi) \le \varepsilon$, then, for $C := \sqrt{\min\{\|\Sigma_{\hat{\pi}}\|_2, \|\Sigma_\pi\|_2\}}$,

$$\max_i |\sigma_{\hat{\pi},i} - \sigma_{\pi,i}| \le 2\varepsilon \ \ \textit{and} \ \ \|\Sigma_{\hat{\pi}} - \Sigma_\pi\|_2 < 3C\varepsilon + 6\varepsilon^2.$$

Remark 3.2. Since $\mathcal{W}_p(\hat{\pi}, \pi)$ is an increasing function of p (which follows from Jensen's inequality), $\mathcal{W}_p(\hat{\pi}, \pi) \le \varepsilon$ for $p \ge 1$ implies $\mathcal{W}_1(\hat{\pi}, \pi) \le \varepsilon$, and for $p \ge 2$ implies $\mathcal{W}_2(\hat{\pi}, \pi) \le \varepsilon$. Thus, $\mathcal{W}_2(\hat{\pi}, \pi) \le \varepsilon$ also suffices for the mean and MAD bounds.

4. Computationally efficient error bounds

In this section we return to the question of computationally efficient posterior error bounds. In particular, although we have shown that the Wasserstein distance provides direct control of the error in approximate posterior summaries of interest, it itself is not tractable to compute or estimate. Our general strategy in this section is use standard variational objectives – namely, $\mathsf{ELBO}(\xi)$ and $\mathsf{CUBO}_\alpha(\xi)$ – to bound the Wasserstein distance. We thereby achieve bounds on the error of posterior summaries by Theorem 3.1. More detail about the intuition and proofs for our results in this section can be found in Section 5 and the Appendix.

Our process consists of two steps. First, we use tail properties of the distribution $\xi \in \mathcal{Q}$ to arrive at bounds on the Wasserstein distance via the KL or α-divergence. Second, we use $\mathsf{ELBO}(\xi)$ and $\mathsf{CUBO}_\alpha(\xi)$ to bound the KL and α-divergences. After describing these steps, we detail our resulting algorithm.

For the first step, we start by defining the moment constants $C_p^{\mathrm{PI}}(\xi)$ and $C_p^{\mathrm{EI}}(\xi)$ and associated tail behaviors. For $p \ge 1$, we say that ξ is *p-polynomially integrable* if

$$C_p^{\mathrm{PI}}(\xi) := 2 \inf_{\theta_0} \left\{ \int \|\theta - \theta_0\|_2^p \xi(\mathrm{d}\theta) \right\}^{\frac{1}{p}} < \infty,$$

and that ξ is *p-exponentially integrable* if

$$C_p^{\mathrm{EI}}(\xi) := 2 \inf_{\theta_0, \epsilon > 0} \left[\frac{1}{\epsilon} \left\{ \frac{3}{2} + \log \int e^{\epsilon \|\theta - \theta_0\|_2^p} \xi(\mathrm{d}\theta) \right\} \right]^{\frac{1}{p}} < \infty.$$

Assuming the variational approximation $\hat{\pi}$ has either exponential (respectively, polynomial) tails, our next result provides a bound on the p-Wasserstein distance using the KL divergence (respectively, the 2-divergence).

Proposition 4.1. *If $\pi \ll \hat{\pi}$,*[3] *then*

$$\mathcal{W}_p(\hat{\pi}, \pi) \le C_{2p}^{\mathrm{PI}}(\hat{\pi}) \left[\exp\left\{ \mathrm{D}_2(\pi \mid \hat{\pi}) \right\} - 1 \right]^{\frac{1}{2p}}$$

and

$$\mathcal{W}_p(\hat{\pi}, \pi) \le C_p^{\mathrm{EI}}(\hat{\pi}) \left[\mathrm{KL}(\pi \mid \hat{\pi})^{\frac{1}{p}} + \{\mathrm{KL}(\pi \mid \hat{\pi})/2\}^{\frac{1}{2p}} \right].$$

For the second step, our next result uses $\mathsf{ELBO}(\xi)$ and $\mathsf{CUBO}_\alpha(\xi)$ to bound the KL and α-divergences. For $\alpha > 1$ and any distribution η, define

$$\mathrm{H}_\alpha(\xi, \eta) := \frac{\alpha}{\alpha - 1} \left\{ \mathsf{CUBO}_\alpha(\xi) - \mathsf{ELBO}(\eta) \right\}.$$

Lemma 4.2. *For any distribution η such that $\pi \ll \eta$,*

$$\mathrm{KL}(\pi \mid \hat{\pi}) \le \mathrm{D}_\alpha(\pi \mid \hat{\pi}) \le \mathrm{H}_\alpha(\hat{\pi}, \eta).$$

Then, combining Proposition 4.1 and Lemma 4.2 yields the desired bounds on the p-Wasserstein distance given only the efficiently computable quantities $C_{2p}^{\mathrm{PI}}(\hat{\pi})$, $C_p^{\mathrm{EI}}(\hat{\pi})$, $\mathsf{CUBO}_\alpha(\hat{\pi})$, and $\mathsf{ELBO}(\eta)$.

Theorem 4.3. *For any $p \ge 1$ and any distribution η, if $\pi \ll \hat{\pi}$, then*

$$\mathcal{W}_p(\hat{\pi}, \pi) \le C_{2p}^{\mathrm{PI}}(\hat{\pi}) \left[\exp\left\{ \mathrm{H}_2(\hat{\pi}, \eta) \right\} - 1 \right]^{\frac{1}{2p}}$$

and

$$\mathcal{W}_p(\hat{\pi}, \pi) \le C_p^{\mathrm{EI}}(\hat{\pi}) \left[\mathrm{H}_2(\hat{\pi}, \eta)^{\frac{1}{p}} + \{\mathrm{H}_2(\hat{\pi}, \eta)/2\}^{\frac{1}{2p}} \right].$$

Finally, our Algorithm 1 details how to use Theorem 4.3 in practice. The algorithm has three subroutines, A–C below. (A) We estimate $\mathsf{CUBO}_2(\hat{\pi})$ via Monte Carlo using samples from $\hat{\pi}$. (B) For *any* distribution η, we estimate $\mathsf{ELBO}(\eta)$, again via Monte Carlo using samples from η. (C) Finally, we bound the moment constants $C_p^{\mathrm{PI}}(\hat{\pi})$ and $C_p^{\mathrm{EI}}(\hat{\pi})$ by fixing any choice of θ_0, ϵ and either sampling from $\hat{\pi}$ or evaluating the expectation exactly, depending on $\mathcal{Q}$. Note that in some cases $C_p^{\mathrm{EI}}(\hat{\pi})$ may be infinite if $\hat{\pi}$ does not have sufficiently light tails. Also, since we typically select the variational family $\mathcal{Q}$, we can do so such that any $\xi \in \mathcal{Q}$ has known tail behavior; this control makes the choice of which bound to use in Theorem 4.3 clear given a particular application.

4.1. Related work

Stein discrepancies. Computable Stein discrepancies form an alternative approach for evaluating variational approximations (Gorham and Mackey, 2015; 2017; Gorham et al.,

[3] $\pi \ll \hat{\pi}$ denotes π is absolutely continuous with respect to $\hat{\pi}$.

Algorithm 1: Bound on $\mathcal{W}_p(\hat{\pi}, \pi)$ via variational objectives

Data: approximate posterior $\hat{\pi}$, distribution η, unnormalized posterior π', Monte Carlo sample size $T \in \mathbb{N}$

$(\theta_{\hat{\pi},s})_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} \hat{\pi}$ /* Obtain samples from $\hat{\pi}$ */

$\widehat{\mathsf{CUBO}_2(\hat{\pi})} \leftarrow \frac{1}{2}\log\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mathrm{d}\pi'}{\mathrm{d}\hat{\pi}}(\theta_{\hat{\pi},s})\right)^2\right\}$ /* Monte Carlo approximation to CUBO */

$(\theta_{\eta,s})_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} \eta$ /* Obtain samples from η */

$\widehat{\mathsf{ELBO}(\eta)} \leftarrow \frac{1}{T}\sum_{t=1}^{T}\log\frac{\mathrm{d}\pi'}{\mathrm{d}\eta}(\theta_{\eta,s})$ /* Monte Carlo approximation to ELBO */

$\widehat{\mathrm{H}_2(\hat{\pi},\eta)} \leftarrow 2\big\{\widehat{\mathsf{CUBO}_2(\hat{\pi})} - \widehat{\mathsf{ELBO}(\eta)}\big\}$ /* compute 2-divergence bound estimate */

// compute bounds on moment constants either analytically or using Monte Carlo

$\widehat{C_p^{\mathrm{EI}}(\hat{\pi})}, \widehat{C_{2p}^{\mathrm{PI}}(\hat{\pi})} \leftarrow \textsc{UpperBounds}(\hat{\pi}, p)$

$B_{\mathrm{E}} \leftarrow \widehat{C_p^{\mathrm{EI}}(\hat{\pi})}\big[\widehat{\mathrm{H}_2(\hat{\pi},\eta)}^{\frac{1}{p}} + \big\{\widehat{\mathrm{H}_2(\hat{\pi},\eta)}/2\big\}^{\frac{1}{2p}}\big]$ /* exponential version of W_p bound */

$B_{\mathrm{P}} \leftarrow \widehat{C_{2p}^{\mathrm{PI}}(\hat{\pi})}\big[\exp\big\{\widehat{\mathrm{H}_2(\hat{\pi},\eta)}\big\} - 1\big]$ /* polynomial version of W_p bound */

return $\min\{B_{\mathrm{P}}, B_{\mathrm{E}}\}$

2019). In particular, the Stein discrepancy between the posterior and variational approximation could be approximated using samples from the variational approximation. However, Stein discrepancy-based bounds on the Wasserstein distance require knowledge about the posterior that is usually not available without additional analytic effort. Moreover, unless the posterior has sub-Gaussian tails, there is not yet an automated, turnkey way to select an appropriate Stein operator that guarantees control of the Wasserstein distance (Erdogdu et al., 2018; Gorham et al., 2019). We expect heaver-than-Gaussian tails in many applications.

Pareto-smoothed importance sampling and $\hat{k}$.

Pareto-smoothed importance sampling (PSIS; Vehtari et al., 2019) is a method for reducing the variance of importance sampling estimators. The key quantity computed in PSIS is $\hat{k}$, which is an estimate of $k := \inf\{k' \mid \mathrm{D}_{1/k'}(\pi \mid \hat{\pi}) < \infty\}$. When $k \le 0.5$, the 2-divergence is finite and hence our efficiently computable Wasserstein distance bounds are finite. Yao et al. (2018) suggest using $\hat{k}$ as a measure of the quality of $\hat{\pi}$. Based on the empirical results and informal arguments of Vehtari et al. (2019), they propose that $\hat{k} \le 0.5$ indicates a good variational approximation and $\hat{k} \in [0.5, 0.7]$ indicates minimal acceptability. In all cases the authors suggest using PSIS to improve estimates of posterior expectations. However, the link between a small $\hat{k}$ value and a high-quality posterior approximation is only heuristic. We find empirically in Section 6 that poor posterior approximations can have small $\hat{k}$ values.

4.2. Importance sampling

For evaluation, we advise using our theoretically sound and efficient Algorithm 1 instead of $\hat{k}$. But we agree with Yao et al. (2018) that it is prudent to improve the accuracy of variational approximations via importance sampling. Namely, if we have samples $\theta_1, \dots, \theta_T \sim \hat{\pi}$, we can define importance weights $w_t := \pi'(\theta_t)/\hat{\pi}(\theta_t)$ and self-normalized weights $\tilde{w}_t := w_t / \sum_{t=1}^{T} w_t$. Then, the importance sampling estimator for $\int \phi \,\mathrm{d}\pi$ is $\sum_{t=1}^{T}\tilde{w}_t\phi(\theta_t)$. Importance sampling can decrease the bias as the cost of some additional variance relative to the simple Monte Carlo estimate $T^{-1}\sum_{t=1}^{T}\phi(\theta_t)$.

Our approach to bounding the Wasserstein distance in terms of the α-divergence has intriguing connections to the theory of importance sampling. As pointed out by Dieng et al. (2017), minimizing the 2-divergence is equivalent to minimizing the variance of the (normalized) importance weight $\pi(\theta_t)/\hat{\pi}(\theta_t)$, which is equal to $\exp\{\mathrm{D}_2(\pi \mid \hat{\pi})\} - 1$. Moreover, the estimation error of importance sampling can be bounded as a function of $\mathrm{KL}(\pi \mid \hat{\pi})$ (Chatterjee and Diaconis, 2018), which is upper bounded by $\mathrm{D}_2(\pi \mid \hat{\pi})$. Thus, minimizing the 2-divergence simultaneously leads to better importance distributions and smaller Wasserstein error – as long as the moments of the variational approximation do not increase disproportionately to the 2-divergence decrease (in practice such pathological behavior appears to be unusual; see Section 6 and Dieng et al. (2017, §3)).

4.3. A workflow for variational inference

Based on Theorem 4.3 and our discussion in Section 4.2, we suggest a number of deviations from the typical variational inference procedure. The usual approach to variational inference is (1) to choose $\mathcal{D}_\pi(\xi) = \mathrm{KL}(\xi \mid \pi)$ (i.e., to minimize $\mathsf{ELBO}(\xi)$), and (2) to use (products of) Gaussians as the variational family $\mathcal{Q}$, which can be optimized using black-box methods (Carpenter et al., 2017a; Kucukelbir et al., 2015; Ranganath et al., 2014; Salvatier et al., 2016).

By contrast, our workflow integrates checks based on our novel bounds. Moreover, we focus on ensuring sufficiently heavy tails in the approximation to capture means and variances in the exact posterior. Along these lines, we recommend minimizing $\mathsf{CUBO}_2(\xi)$ rather than $\mathsf{ELBO}(\xi)$ to start. In addition, we advise against using Gaussian variational families and instead suggest using a product of t-distributions in black-box methods. If the posterior distribution has polynomial tails, those will dictate appropriate t-distribution degrees of freedom. Otherwise, the degrees of freedom can be set to a large value such as 40 or 100. We show the following workflow in action in Section 6.

1. If feasible, select a variational family $\mathcal{Q}$ with sufficiently heavy tails that $k \leq 0.5$ (that is, such that $\mathrm{D}_2(\pi \mid \xi) < \infty$ for all $\xi \in \mathcal{Q}$).
2. Use CHIVI to choose a variational approximation $\hat{\pi} \in \mathcal{Q}$ that minimizes $\mathsf{CUBO}_2(\xi)$.
3. If there is no guarantee that $k \leq 0.5$ (see step 1), compute $\hat{k}$. If $\hat{k} > 0.5$, then refine the choice of $\mathcal{Q}$ or reparameterize the model to make it more conducive to approximation by distributions in $\mathcal{Q}$.
4. Compute $\mathsf{ELBO}(\hat{\pi})$ and $\mathsf{CUBO}_2(\hat{\pi})$.
5. Optionally, further optimize $\mathsf{ELBO}(\xi)$ to obtain a tighter lower bound on the marginal likelihood.
6. Use Lemma 4.2 to compute a bound $\bar{\delta}_2 \geq \mathrm{D}_2(\pi \mid \hat{\pi})$
7. Use Theorem 4.3 to compute a bound $\bar{w}_2 \geq \mathcal{W}_2(\pi, \hat{\pi})$.
8. If $\bar{\delta}_2$ and $\bar{w}_2$ are large, then refine the choice of $\mathcal{Q}$ or reparameterize the model to make it more conducive to approximation by distributions in $\mathcal{Q}$.
9. If $\bar{\delta}_2$ is small but $\bar{w}_2$ is large, use either standard importance sampling or PSIS to refine the posterior expectations produced by $\hat{\pi}$.
10. If both bounds are small, $\hat{\pi}$ can be used directly to approximate π.

Remark 4.4. What qualifies as a sufficiently small $\bar{w}_2$ will depend on the desired accuracy and natural scale of the problem. $\bar{\delta}_2$ has a more universal scale; in particular, choosing $\bar{\delta}_2 < \log(2) \approx 0.7$ guarantees that $\exp\{\mathrm{D}_2(\hat{\pi} \mid \pi))\} - 1 < 1$. Therefore the variance of the importance weights and the term multiplying $C_p^{\mathrm{PI}}(\xi)$ in Proposition 4.1 and Theorem 4.3 will be less than 1.

5. Transport–divergence inequalities

Next, we develop a deeper understanding of our bound in Theorem 4.3 and its proof. In particular, we explore how the ELBO and CUBO from variational inference can be used to bound the Wasserstein distance, despite one depending on a metric and the other not; we show that our new theory – including variations on our main bound Theorem 4.3 – avoids the strong tail assumptions of existing related work. And we demonstrate that our bound avoids the pathological behavior of the KL divergence from Examples 3.1 and 3.2. Our results in this section are of independent interest beyond Bayesian inference, so we use the notation η and ν to represent two arbitrary distributions; in the Bayesian setting, we would choose $\pi = \eta$ and $\hat{\pi} = \nu$.

Our first challenge is bounding a scale-dependent distance (Wasserstein) with a scale-invariant divergence (KL or α-divergence). To see the scale-invariance, we note a broader result: these divergences are invariant to reparameterization. For a transformation $T : \mathbb{R}^d \to \mathbb{R}^d$, let $T\#\eta$ denote the pushforward measure of η, which is the distribution of the random variable $T(\vartheta)$ for $\vartheta \sim \eta$.

Lemma 5.1. *The KL and α-divergence are invariant under a smooth, invertible transformation T, i.e.,* $\mathrm{D}_\alpha(\eta \mid \nu) = \mathrm{D}_\alpha(T\#\eta \mid T\#\nu)$ *and* $\mathrm{KL}(\eta \mid \nu) = \mathrm{KL}(T\#\eta \mid T\#\nu)$.

Our proof is in Appendix D.7; Yao et al. (2018) make a similar observation. A simple example illustrates that Wasserstein is not invariant to reparameterization. For $\sigma \in \mathbb{R}_+$, let $\nu_\sigma(\theta) := \sigma^{-d}\nu(\theta/\sigma)$ define the rescaled version of ν (with η_σ defined analogously). Then $\mathcal{W}_p(\eta_\sigma, \nu_\sigma) = \sigma\mathcal{W}_p(\eta, \nu)$. It follows that any bound of a scale-dependent distance such as Wasserstein using a scale-invariant divergence must incorporate some notion of scale.

There are a number of existing bounds on $\mathcal{W}_p(\eta, \nu)$ via $\mathrm{KL}(\eta \mid \nu)$, generally referred to as *transport–entropy inequalities* (with reference to the other name for KL divergence, *relative entropy*). As just discussed, these require a scale parameter to modulate the bound. Existing bounds, however, are not sufficient for our present purposes since they typically require impractically strong tail assumptions. In particular, Theorem B.3 in Appendix B, due to Bobkov and Götze (1999); Djellout et al. (2004), requires that ν be 2-exponentially integrable and hence have lighter or equal tails to a Gaussian. The following theorem – which we use in Theorem 4.3 – requires only exponential tails to bound the 1-Wasserstein distance.

Proposition 5.2 (Bolley and Villani (2005, Corollary 2.3))**.** *Assume ν is p-exponentially integrable for some $p \geq 1$. Then for all $\eta \ll \nu$,*

$$\mathcal{W}_p(\eta, \nu) \leq C_p^{\mathrm{EI}}(\nu) \left[\mathrm{KL}(\eta \mid \nu)^{\frac{1}{p}} + \{\mathrm{KL}(\eta \mid \nu)/2\}^{\frac{1}{2p}}\right].$$

However, many posteriors of interest have much heavier tails – often with at most polynomial decay. For example, neither inverse Gamma distributions nor t-distributions

	centered df = 40		non-centered df = 40		non-centered df = 10	
	KLVI	CHIVI	KLVI	CHIVI	KLVI	CHIVI
D_2 bound	17	11	3.3	1.6	2.9	6.6
$\hat{k}$	0.90	0.87	0.69	0.53	0.52	0.42
W_2 bound	1600	490	110	190	86	2900
mean error	0.91	0.80	1.5	0.43	1.3	1.1
with PSIS	0.93	1.17	0.07	0.06	0.07	0.07
covariance error	10.0	8.4	3.7	10.9	3.9	38
with PSIS	6.8	4.6	1.1	1.0	1.0	1.0

Table 1: Results for eight schools model for the parameter vector $(\mu, \log \tau, \theta_1, \ldots, \theta_8)$. The mean and covariance errors are defined as, respectively, $\|m_\pi - m_{\hat{\pi}}\|_2$ and $\|\Sigma_\pi - \Sigma_{\hat{\pi}}\|_2^{1/2}$. We use the square root for the covariance error in order to place it on the same scale as the mean error and the 2-Wasserstein bound.

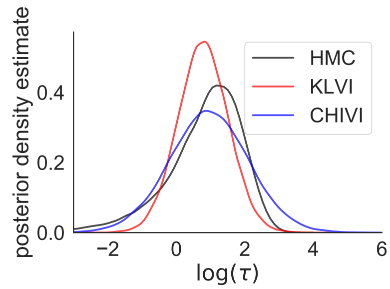


Figure 1: Posterior approximations to $\log \tau$ using HMC (ground truth), KLVI, and CHIVI with the non-centered parameterization. The right tail of the CHIVI approximation is too heavy, leading to overestimation of the variance of θ.

with $h < \infty$ degrees of freedom have exponential tails. Moreover, if we wish to bound the 2-Wasserstein distance, Proposition 5.2 requires the problematic Gaussian tails assumption.

In contrast to these past results, our work provides bounds on Wasserstein distances assuming only polynomial tail decay. We achieve these bounds by incorporating more general α-divergences; we call these new bounds *transport–divergence inequalities*. For example, Proposition 5.3 is a particularly simple bound on the p-Wasserstein distance in terms of just the 2-divergence when ν has finite $(2p)$th moment. We use this result, together with Lemma 4.2 and Proposition 5.2, to prove Theorem 4.3 above.

Proposition 5.3. *Assume ν is $2p$-polynomially integrable for some $p \geq 1$. Then for all $\eta \ll \nu$,*

$$\mathcal{W}_p(\eta, \nu) \leq C_{2p}^{\mathrm{PI}}(\nu) \left[\exp\left\{\mathrm{D}_2(\eta \mid \nu)\right\} - 1\right]^{\frac{1}{2p}}.$$

In Appendix C, we show how to achieve tighter bounds than Proposition 5.3; these can be combined with Lemma 4.2 to arrive at results like Theorem 4.3, at the price of additional complexity in the statements of the bounds.

Finally, we check that, even though our transport inequalities use KL and α-divergences, our bounds do not suffer the pathologies in Examples 5.1 and 5.2; rather, our bounds capture the growth in error, as desired. We provide complete details for the examples in Appendix C.

Example 5.1 (cf. Examples 3.1 and 3.3)**.** For a fixed $k \in (0, \infty)$, let $\eta = \mathsf{Weibull}(k/2, 1)$ and $\nu = \mathsf{Weibull}(k, 1)$. Then, for $\alpha > 1$, $\mathrm{D}_\alpha(\eta \mid \nu) = \infty$. On the other hand, $\mathrm{D}_\alpha(\nu \mid \eta) < \infty$; but, as $k \searrow 0$, the moment constant from Proposition 5.3 satisfies $C_p^{\mathrm{PI}}(\eta) \nearrow \infty$. □

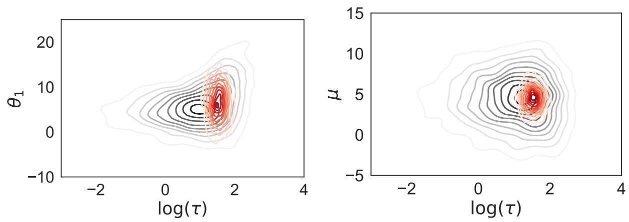
Example 5.2 (cf. Example 3.2)**.** If η is a standard normal measure and $\nu = \mathcal{T}_h$ is a standard t-distribution with $h \geq 2$ degrees of freedom, then $\mathrm{D}_\alpha(\eta \mid \nu) < \infty$. However, as $h \searrow 2$, we have $C_p^{\mathrm{PI}}(\nu) \nearrow \infty$. □

6. Case study: the eight schools model

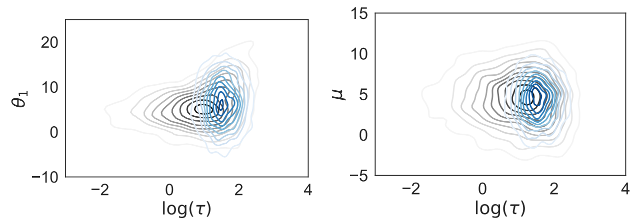
Next we demonstrate our variational inference workflow and the usefulness of our bounds. In particular, we apply variational inference to approximate the posterior for the eight schools data and model (Gelman et al., 2013, Sec. 5.5), a canonical example of a Bayesian hierarchical analysis. Yao et al. (2018) previously considered this model in the setting of evaluating variational inference. In the eight schools data, we have observations corresponding to the mean y_n and standard deviation σ_n of a treatment effect at each of eight schools, indexed by $n \in \{1, \ldots, 8\}$. The goal is to estimate the overall treatment effect μ, the standard deviation τ of school-level treatment effects, and the true school-level treatment effects θ_n:

$$\begin{aligned} y_n \mid \theta_n &\sim \mathcal{N}(\theta_n, \sigma_n), \;\; \theta_n \mid \mu, \tau \sim \mathcal{N}(\mu, \tau), \\ \mu &\sim \mathcal{N}(0, 5), \qquad\qquad \tau \sim \mathsf{half\text{-}Cauchy}(0, 5). \end{aligned} \tag{1}$$

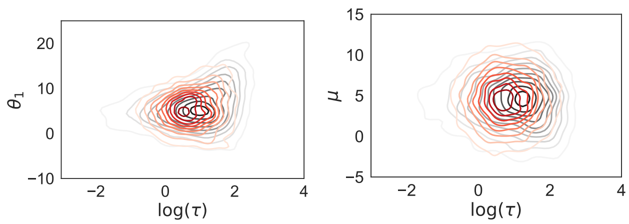
We chose t-distributions with 40 degrees of freedom as the variational family $\mathcal{Q}$ in order to guarantee the 2-divergence would be finite for any $\xi \in \mathcal{Q}$ (step 1 of the workflow). We used black-box variational inference, minimizing each of the KL divergence (KLVI) and the 2-divergence (CHIVI), respectively. While we advocate using CHIVI in general (step 2), we also used KLVI here for illustration. For now we defer discussion of the $\hat{k}$ values since by construction we know that $k < \infty$. Following steps 4–7 of the workflow, we used the CUBO and ELBO to compute the 2-divergence



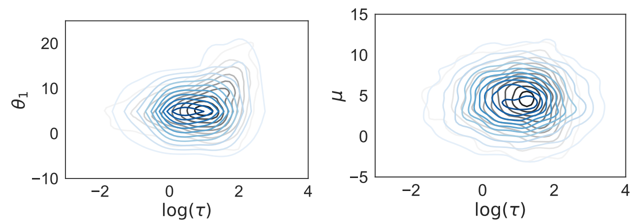
(a) KLVI, centered



(b) CHIVI, centered



(c) KLVI, non-centered



(d) CHIVI, non-centered

Figure 2: Approximate posteriors for variational approximations (KLVI red, CHIVI blue) and for HMC (black; truth).

and 2-Wasserstein distance bounds. These results appear in the first two columns of Table 1 labeled "centered / df = 40". The standard deviation of the y_n is 9.8 and the median σ_n value is 11, which suggests the overall scale of the problem is roughly 10. Both the KLVI and CHIVI approximations had large 2-divergence bounds (greater than 10) and very large 2-Wasserstein bounds (greater than 400), which are reflected in the large mean and standard deviation errors. Thus, the 2-divergence and 2-Wasserstein bounds accurately reflect the poor quality of the variational approximations. Following guidance of step 8, we need to either reparameterize the model or choose a different $\mathcal{Q}$.

Fig. 2(a,b) compares approximate posteriors from KLVI, CHIVI, and Hamiltonian Monte Carlo (HMC; Neal, 2011) – namely the No-U-Turn sampler (Hoffman and Gelman, 2014) in Stan (Carpenter et al., 2017b). The HMC samples serve as ground truth. This comparison illustrates why the eight schools model in Eq. (1) is not conducive to variational inference when $\xi \in \mathcal{Q}$ is a product distribution: the conditional variance of any θ_n is strongly dependent on τ.

To improve the variational approximation, we can instead use the *non-centered* parameterization, which decouples θ and τ through the transformation $\tilde{\theta}_n = (\theta_n - \mu_n)/\tau$:

$$y_n \mid \tilde{\theta}_n \sim \mathcal{N}(\mu + \tau\tilde{\theta}_n), \qquad \tilde{\theta}_n \sim \mathcal{N}(0, 1).$$

Repeating steps 2–7 of the workflow, the results of using KLVI and CHIVI appear in the two columns of Table 1 labeled "non-centered / df = 40". The 2-divergence and 2-Wasserstein bounds for the non-centered case are substantially smaller than in the centered case. Examining Fig. 2(c,d), which shows comparisons of the posteriors estimated using KLVI and CHIVI with the non-centered parameterization, we can see the generally superior quality of the non-centered approximations. Notably, although the CHIVI approximation has a smaller 2-divergence bound than KLVI here, it has a larger 2-Wasserstein bound than KLVI. This larger bound accurately reflects that CHIVI provides an inferior approximation to the standard deviation. The problem, as shown in Fig. 1, is that the CHIVI approximation greatly overestimates the variance of θ due to the right tail of the distribution of $\log\tau$ being too heavy.

With the non-centered parameterization, the 2-divergence and 2-Wasserstein bounds are too large to provide confidence in using either the KLVI or CHIVI approximation directly. But the 2-divergence bounds are sufficiently small to suggest that an importance sampling correction should substantially improve estimation accuracy, as suggested by workflow step 9. The errors in the PSIS-based estimates in Table 1 show that importance sampling offers an improvement for the non-centered case. The CHIVI approximation providing the best importance distribution, in agreement with the smaller 2-divergence bound. We also applied PSIS to samples from the centered variational approximations. While PSIS generally improved the estimate, for the centered CHIVI approximation the mean estimate was worse, which is in line with expectations given the larger 2-divergence bound. Recall that the variance of the vanilla importance weights is $\exp\{\mathrm{D}_2(\pi \mid \xi)\} - 1$.

Using $\hat{k}$ to evaluate $\hat{\pi}$. So far, as advocated in our proposed workflow, we have used PSIS for improving posterior inference. Now we turn to evaluating the use of $\hat{k}$ as a diagnostic for the approximation quality of $\hat{\pi}$. Clearly the $\hat{k}$ values (shown in Table 1) are somewhat misleading (or at least pessimistic) insofar as they are greater than 0.5

even though by construction of $\mathcal{Q}$ we know that $k < 0.5$. Moreover, it is not clear from these results whether the cutoff of 0.5 for $\hat{k}$ is in fact the correct diagnostic for determining whether $\hat{\pi}$ is a good approximation to π. To better understand the behavior of $\hat{k}$ relative to our bounds on the 2-divergence and 2-Wasserstein distance, we ran KLVI and CHIVI for the non-centered model, but this time we used t-distributions with 10 degrees of freedom as the variational family $\mathcal{Q}$. Heavier tails on the variational approximation should decrease k and $\hat{k}$ – since the importance weights will have more finite moments. The results appear in the two columns of Table 1 labeled "non-centered / df = 10". While these approximations produced the smallest $\hat{k}$ values, with CHIVI having the smallest and the only $\hat{k} < 0.5$, the CHIVI approximation was in fact quite poor, with large mean error and extremely large covariance error. The poor quality of the CHIVI approximation was, however, accurately reflected in a somewhat larger 2-divergence bound and a very large 2-Wasserstein distance bound.

In sum, while $\hat{k}$ does provide a useful diagnostic for when $\hat{\pi}$ will serve as a good importance distribution, it does not provide a reliable heuristic for the accuracy of $\hat{\pi}$ as an approximation to π. On the other hand, our 2-divergence and 2-Wasserstein distance bounds together offer a more accurate and theoretically sound diagnostic tool for variational approximations. And, as we have shown through both theory and experiment, our variational inference workflow potentially provides a framework for making variational methods more competitive with Markov chain Monte Carlo. We conclude by noting that our work complements recent proposals for making variational approximations arbitrarily accurate (Campbell and Li, 2019; Guo et al., 2016; Locatello et al., 2018a;b; Miller et al., 2017; Wang, 2016) since our bounds can provide a stopping criteria for when a variational approximation no longer needs to be improved.

Acknowledgements

The authors thank Daniel Simpson, Lester Mackey, Arthur Gretton, and Pierre Jacob for valuable discussions and many useful references. M. Kasprzak was supported in part by an EPSRC studentship and FNR grant FoRGES (R-AGR-3376-10). T. Campbell was supported by a National Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and Discovery Launch Supplement. T. Broderick was supported in part by an NSF CAREER Award, an ARO YIP Award, the Office of Naval Research, a Sloan Research Fellowship, the CSAIL-MSR Trustworthy AI Initiative, and DARPA.

References

P. Alquier and J. Ridgway. Concentration of tempered posteriors and of their variational approximations. *arXiv.org*, June 2017.

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17:1–41, 2016.

C. Bauckhage. Computing the Kullback-Leibler Divergence between two Weibull Distributions. *arXiv.org*, 2013.

C. M. Bishop. *Pattern recognition and machine learning*, chapter 10: Approximate Inference. Springer, 2006.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

S. G. Bobkov and F. Götze. Exponential Integrability and Transportation Cost Related to Logarithmic Sobolev Inequalities. *Journal of Functional Analysis*, 163:1–28, 1999.

F. Bolley and C. Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la faculte des sciences de Toulouse*, 13(3):331–352, 2005.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

T. D. Bui, J. Yan, and R. E. Turner. A Unifying Framework for Gaussian Process Pseudo-Point Approximations using Power Expectation Propagation. *Journal of Machine Learning Research*, 18:1–72, Oct. 2017.

T. Campbell and X. Li. Universal boosting variational inference. In *Advances in Neural Information Processing Systems*, 2019.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), 2017a.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017b.

S. Chatterjee and P. Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, Apr. 2018.

B.-E. Chérief-Abdellatif and P. Alquier. Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12 (2):2995–3035, 2018.

A. Cichocki and S.-I. Amari. Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities. *Entropy*, 12(6):1532–1568, June 2010.

S. Claici, E. Chien, and J. Solomon. Stochastic Wasserstein barycenters. In *International Conference on Machine Learning*, 2018.

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

A. B. Dieng, D. Tran, R. Ranganath, J. Paisley, and D. M. Blei. Variational Inference via χ Upper Bound Minimization. In *Advances in Neural Information Processing Systems*, 2017.

H. Djellout, A. Guillin, and L. Wu. Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *The Annals of Probability*, 32(3B):2702–2732, July 2004.

A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, Nov. 2019.

A. Durmus, S. Majewski, and B. Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20:1–46, 2019.

A. Eberle and M. B. Majka. Quantitative contraction rates for Markov chains on general state spaces. *Electronic Journal of Probability*, 24(0):1–36, 2019.

M. A. Erdogdu, L. Mackey, and O. Shamir. Global Nonconvex Optimization with Discretized Diffusions. In *Advances in Neural Information Processing Systems*, 2018.

A. Gelman, J. Carlin, H. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, third edition, 2013.

J. Gorham and L. Mackey. Measuring Sample Quality with Stein's Method. In *Advances in Neural Information Processing Systems*, 2015.

J. Gorham and L. Mackey. Measuring Sample Quality with Kernels. In *International Conference on Machine Learning*, 2017.

J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928, Oct. 2019.

N. Gozlan. A characterization of dimension free concentration in terms of transportation inequalities. *The Annals of Probability*, 37(6):2480–2498, Nov. 2009.

F. Guo, X. Wang, K. Fan, T. Broderick, and D. Dunson. Boosting variational inference. In *Advances in Neural Information Processing Systems*, 2016.

J. M. Hernández-Lobato, Y. Li, M. Rowland, T. D. Bui, D. Hernández-Lobato, and R. E. Turner. Black-Box Alpha Divergence Minimization. In *International Conference on Machine Learning*, 2016.

M. D. Hoffman and A. Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

A. Joulin and Y. Ollivier. Curvature, concentration and error estimates for Markov chain Monte Carlo. *The Annals of Probability*, 38(6):2418–2442, 2010.

A. Kucukelbir, R. Ranganath, A. Gelman, and D. M. Blei. Automatic Variational Inference in Stan. In *Advances in Neural Information Processing Systems*, June 2015.

Y. Li and R. E. Turner. Rényi Divergence Variational Inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

F. Locatello, G. Dresdner, R. Khanna, I. Valera, and G. Rätsch. Boosting black box variational inference. In *Advances in Neural Information Processing Systems*, 2018a.

F. Locatello, R. Khanna, J. Ghosh, and G. Rätsch. Boosting variational inference: an optimization perspective. In *International Conference on Artificial Intelligence and Statistics*, 2018b.

N. Madras and D. Sezer. Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances. *Bernoulli*, 16(3):882–908, Aug. 2010.

A. Miller, N. Foti, and R. Adams. Variational boosting: iteratively refining posterior approximations. In *International Conference on Machine Learning*, 2017.

R. M. Neal. Mcmc using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

D. Pati, A. Bhattacharya, and Y. Yang. On Statistical Optimality of Variational Bayes. *AISTATS*, 2018.

R. Ranganath, S. Gerrish, and D. M. Blei. Black Box Variational Inference. In *International Conference on Artificial Intelligence and Statistics*, pages 814–822, 2014.

C. P. Robert. *The Bayesian Choice*. Springer, New York, NY, 1994.

D. Rudolf and N. Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 4A: 2610–2639, 2018.

J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

S. Srivastava, C. Li, and D. B. Dunson. Scalable Bayes via barycenter in Wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018.

A. Vehtari, D. Simpson, A. Gelman, Y. Yao, and J. Gabry. Pareto Smoothed Importance Sampling. *arXiv.org*, July 2019.

C. Villani. *Topics in Optimal Transportation*. American Mathematical Soc., 2003.

C. Villani. *Optimal transport: old and new*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh. (Non-) asymptotic properties of Stochastic Gradient Langevin Dynamics. *Journal of Machine Learning Research*, 17 (159):1–48, 2016.

M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

X. Wang. Boosting variational inference: theory and examples. Master's thesis, Duke University, 2016.

Y. Wang and D. M. Blei. Frequentist Consistency of Variational Bayes. *Journal of the American Statistical Association*, 17(239):1–86, June 2018.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Yes, but Did It Work?: Evaluating Variational Inference. In *International Conference on Machine Learning*, 2018.

A. Further details on Examples 3.1 to 3.3

For Example 3.1(A), we let $\hat{\pi} = \mathsf{Weibull}(k/2, 1)$ and $\pi = \mathsf{Weibull}(k, 1)$. Let γ be the Euler-Mascheroni constant and Γ be the gamma function. We obtain (Bauckhage, 2013)

$$\mathrm{KL}(\hat{\pi} \mid \pi) = -\log(2) + \gamma + \Gamma(3) - 1 < 0.9.$$

Using the well-known formulas for the mean and variance of the Weibull distribution, we have $m_{\hat{\pi}} = \Gamma(1+2/k)$, $m_\pi = \Gamma(1+1/k)$, and $\sigma_\pi^2 = \Gamma(1+2/k) - \{\Gamma(1+1/k)\}^2$. Hence, $\lim_{k\searrow 0}(m_{\hat{\pi}} - m_\pi)^2/\sigma_\pi^2 = \infty$.

For Example 3.1(B), let $\hat{\pi} = \mathsf{Weibull}(k, 1)$ and $\pi = \mathsf{Weibull}(k/2, 1)$. We obtain

$$\mathrm{KL}(\hat{\pi} \mid \pi) = \log(2) - \gamma/2 + \Gamma(3/2) - 1 < 0.3.$$

By the same argument as above, $\lim_{k\searrow 0}(m_{\hat{\pi}} - m_\pi)^2/\sigma_{\hat{\pi}}^2 = \infty$.

For Example 3.2, using Jensen's inequality, it is straightforward to show that

$$\begin{aligned}\mathrm{KL}(\hat{\pi} \mid \pi) &= \log[\Gamma(h/2)h^{1/2}/\Gamma\{(h+1)/2\}] - 0.5\log(2e) + 0.5(h+1)\mathbb{E}\left\{\log\left(1 + \vartheta^2/h\right)\right\} \\ &\leq \log[\Gamma(h/2)h^{1/2}/\Gamma\{(h+1)/2\}] - 0.5\log(2e) + 0.5(h+1)\log\left\{1 + \mathbb{E}(\vartheta^2)/h\right\} \\ &\leq \log[\Gamma(h/2)h^{1/2}/\Gamma\{(h+1)/2\}] - 0.5\log(2e) + 0.5(h+1)\log\left\{1 + 1/h\right\}.\end{aligned}$$

For $h = 2$, the bound is less than 0.23. A straightforward calculation shows that $\mathrm{dKL}(\hat{\pi} \mid \mathcal{T}_h)/\mathrm{d}h < 0$ for all $h \geq 2$. Thus, $\mathrm{KL}(\hat{\pi} \mid \mathcal{T}_h) < 0.23$ for all $h \geq 2$.

Finally, we observe $\lim_{h\searrow 2}\sigma_{\mathcal{T}_h}^2 = \infty$.

For Example 3.3, we choose $\pi = \mathsf{Weibull}(k, 1)$ and $\hat{\pi} = \mathsf{Weibull}(k/2, 1)$ for $k > 0$. Note that $\lim_{k\downarrow 0}\frac{\sigma_{\hat{\pi}}^2}{\sigma_\pi^2} = \infty$ and $\lim_{k\downarrow 0}(m_{\hat{\pi}} - m_\pi)^2/\sigma_\pi^2 = \infty$. On the other hand, letting $f_{\hat{\pi}}$ and f_π be the densities of $\hat{\pi}$ and π, respectively, we have

$$\begin{aligned}&\int_0^\infty (f_\pi(x))^2 (f_{\hat{\pi}}(x))^{-1}\, dx \\ &= 2k\int_0^\infty x^{3k/2-1}\exp\left(-2x^k + x^{k/2}\right) dx \\ &\overset{y=x^{k/2}}{=} 4\int_0^\infty y^2\exp\left(-2y^2 + y\right) dy < 1.47752\end{aligned}$$

and so

$$\mathrm{D}_2(\pi \mid \hat{\pi}) = \log\int_0^\infty (f_\pi(x))^2 (f_{\hat{\pi}}(x))^{-1}\, dx < 0.391.$$

Therefore, for any $t > 0$, there exist two distributions $\hat{\pi}$ and π with $\mathrm{D}_2(\pi \mid \hat{\pi})$ bounded by 0.391 yet such that $\sigma_{\hat{\pi}}^2 \geq t\sigma_\pi^2$ and $(m_{\hat{\pi}} - m_\pi)^2 \geq t\sigma_\pi^2$.

B. Transportation–entropy inequality results

Classical transportation–entropy inequalities take the following form.

Definition B.1. For $p \geq 1$ and $\rho > 0$, the distribution ν satisfies a *p-transportation–entropy (or p-Talagrand) inequality with constant* ρ (denoted $\nu \in \mathrm{W}_p\mathrm{H}(\rho)$) if for all $\eta \ll \nu$,

$$\mathcal{W}_p(\eta, \nu) \leq \left\{\frac{2\,\mathrm{KL}(\eta \mid \nu)}{\rho}\right\}^{1/2}.$$

When $p = 1$ there are interpretable necessary and sufficient conditions for $\nu \in \mathrm{W}_1\mathrm{H}(\rho)$. The most important is the p-exponential integrability condition, which we denote by $\nu \in \mathrm{EI}_p(\epsilon)$:

Definition B.2 (cf. Section 4)**.** For $p \geq 1$ and $\epsilon > 0$ the distribution ν is p-exponentially integrable with parameter ϵ (denoted $\nu \in \mathrm{EI}_p(\epsilon)$) if

$$\inf_{\theta_0} \left[\int e^{\epsilon \|\theta - \theta_0\|_2^p} \nu(\mathrm{d}\theta) \right] < \infty.$$

In particular, the following result shows that ν satisfies a 1-transportation–entropy inequality if and only if it has Gaussian tails. Moreover, the ϵ parameter in the corresponding 2-exponential integrability condition essentially determines the precision of the transportation–entropy inequality.

Theorem B.3 (Bobkov and Götze (1999, Theorem 3.1), Djellout et al. (2004, Theorem 2.3))**.** *The following conditions are equivalent:*

1. *For some $\rho > 0$, $\nu \in \mathrm{W}_1\mathrm{H}(\rho)$, as defined in Definition B.1.*
2. *For some $\epsilon > 0$, $\nu \in \mathrm{EI}_2(\epsilon)$, as defined in Definition B.2.*
3. *There exists a constant $c > 0$ such that for every $\phi : \mathbb{R}^d \to \mathbb{R}$ with $\|\phi\|_L \leq 1$ (where $\|\cdot\|_L$ denotes the Lipschitz constant) and every $t \in \mathbb{R}$,*

$$\nu(e^{t\phi}) \leq e^{ct^2}.$$

Moreover, we may take $c = \rho^{-1}$ and

$$c \leq \frac{2}{\epsilon} \sup_{k \geq 1} \left\{ \frac{(k!)^2}{(2k)!} \int \int e^{\epsilon \, \|\theta - \theta'\|_2^2} \nu(\mathrm{d}\theta)\nu(\mathrm{d}\theta') \right\}^{1/k}.$$

Remark B.4. Let $\vartheta \sim \nu$ and for a Lipschitz function $\phi : \mathbb{R}^d \to \mathbb{R}$, let $c_\phi := 2\|\phi\|_L^2 c$. Condition (3) implies that the random variable $\phi(\vartheta)$ is c_ϕ-sub-Gaussian (Boucheron et al., 2013, §2.3). In particular, we have the concentration inequality

$$\mathbb{P}\{\phi(\vartheta) - \nu(\phi) > t\} \leq e^{-\frac{t^2}{2c_\phi}}.$$

The implication (2) $\implies$ (1) from Theorem B.3 can be generalized to cover $p > 1$.

Definition B.5. For $p \geq 1$, the *optimal p-exponential integrability constant* is given by

$$\mathrm{EI}_{2p}^*(\nu, \epsilon) := \inf_{\theta'} \log \int e^{\epsilon \|\theta - \theta'\|_2^p} \nu(\mathrm{d}\theta).$$

Proposition B.6 (Bolley and Villani (2005, Corollary 2.4))**.** *Assume $\nu \in \mathrm{EI}_{2p}(\epsilon)$ (Definition B.2) for some $p \geq 1$ and $\epsilon > 0$ and let*

$$C := 2 \inf_{\epsilon > 0} \left[\frac{1}{2\epsilon} \left\{ 1 + \mathrm{EI}_{2p}^*(\nu, \epsilon) \right\} \right]^{\frac{1}{2p}} < \infty,$$

for $\mathrm{EI}_{2p}^*(\nu, \epsilon)$ defined in Definition B.5. Then for all $\eta \ll \nu$,

$$\mathcal{W}_p(\eta, \nu) \leq C \, \mathrm{KL}(\eta \mid \nu)^{\frac{1}{2p}}.$$

If one can establish that $\nu \in \mathrm{W}_p\mathrm{H}(\rho)$, then the pushforward measure under a Lipschitz transformation also satisfies a p-transportation–entropy inequality.

Lemma B.7. *Assume that for some $\rho > 0$, $\nu \in \mathrm{W}_p\mathrm{H}(\rho)$, and that $\Psi : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz; i.e.,*

$$\|\Psi(\theta) - \Psi(\theta')\|_2 \leq L \, \|\theta - \theta'\|_2 \quad \theta, \theta' \in \mathbb{R}^d.$$

Then $\Psi \# \nu \in \mathrm{W}_p\mathrm{H}(\rho/L^2)$.

We close with the interesting connection that $\nu \in \mathrm{W}_2\mathrm{H}(\rho)$ is equivalent to ν satisfying a dimension-free Gaussian concentration inequality (cf. Remark B.4). While the concentration condition is not necessarily easy to check, it does offer insight into what it means for $\nu \in \mathrm{W}_2\mathrm{H}(\rho)$.

Theorem B.8 (Gozlan (2009, Theorem 1.3)). *For a set $A \subseteq (\mathbb{R}^d)^n$, let $A^t := \{\theta \in (\mathbb{R}^d)^n \mid \exists \theta' \in A : \sum_{i=1}^n \|\theta_i - \theta_i'\|_2^2 \le t^2\}$. The following conditions are equivalent:*

1. *For some $\rho > 0$, $\nu \in \mathrm{W_2H}(\rho)$.*
2. *There exist $a > 0, b > 0$ such that for all $n \in \mathbb{N}$ and measurable $A \subseteq E^n$, with $\nu^{\otimes n}(A) \ge 1/2$, the probability measure $\nu^{\otimes n}$ satisfies*

$$\nu^{\otimes n}(A^t) \ge 1 - be^{-at^2}.$$

C. Our novel transportation–divergence inequality results

We begin by describing two additional novel transportation–entropy inequalities. By combining either of these results Lemma 4.2, we obtain alternative efficiently computable Wasserstein distance bound.

Our first result offers a better dependence on moments in the exponential tails case by using both KL divergence and α-divergence (cf. just KL divergence in Proposition 5.2); however, the bound is more complex than Proposition 5.2. In particular, if ν has exponential tails and we can bound the α-divergence for any $\alpha > 1$, then we can bound the 2-Wasserstein distance.

Theorem C.1. *Assume $\nu \in \mathrm{EI}_{p/2}(\epsilon)$ (Definition B.2) for some $p \ge 1$ and $\epsilon > 0$ and let $\mathrm{EI}_p^*(\nu, \epsilon)$ be defined as in Definition B.5. Let*

$$C(\alpha, \eta, \nu) := \inf_{\epsilon > 0} \left\{ \frac{6}{\epsilon^2} \left[\left(\frac{3\alpha}{\alpha - 1} \right)^2 + 6 + 2\mathrm{EI}_{p/2}^*(\nu, \epsilon)^2 + \mathrm{D}_\alpha(\eta \mid \nu)^2 \right] \right\}^{1/p}.$$

Then for $\alpha > 1$ and $\eta \ll \nu$,

$$\mathcal{W}_p(\eta, \nu) \le C(\alpha, \eta, \nu) \mathrm{KL}(\eta \mid \nu)^{\frac{1}{2p}}.$$

Our second only requires ν to have a finite $(2pq)$th moment in order to bound the p-Wasserstein distance by the relative entropy and the α-divergence, where $q = q(\alpha) := \alpha/(\alpha - 1)$ is the conjugate exponent for α. Thus, it has a higher moment dependence than our Proposition 5.3, but it uses the α-divergence with $\alpha < 2$ (cf. $\alpha = 2$ in Proposition 5.3) and thereby could produce tighter bounds.

Theorem C.2. *Fix $p \ge 1$ and $\alpha > 1$, and let $q = q(\alpha) := \alpha/(\alpha - 1)$. Assume that ν is $2pq$-polynomially integrable, as defined in Section 4, and let*

$$C(\alpha, \eta) := \inf_{\theta'} \left[\left(\int m(\theta', \theta)^{2p} \nu(\mathrm{d}\theta) \right)^{1/2} + \left(\frac{1}{2^{2q-2} q} \int \|\theta' - \theta\|_2^{2pq} \nu(\mathrm{d}\theta) + \frac{4e^{(\alpha-1)\mathrm{D}_\alpha(\eta|\nu)}}{\alpha} \right)^{1/2} \right]^{1/p}.$$

Then for all $\eta \ll \nu$,

$$\mathcal{W}_p(\eta, \nu) \le 2C(\alpha, \eta)\, \mathrm{KL}(\eta \mid \nu)^{\frac{1}{2p}}.$$

Finally, we detail the proofs of our examples, which demonstrate that our bounds do not face the pathologies of KL. First, we work out the details of Example 5.1.

Proof of Example 5.1. Let $\alpha > 1$. Then η and ν have the following densities w.r.t. the Lebesgue measure

$$f_\eta(x) = \frac{k}{2} x^{k/2-1} e^{-x^{k/2}} \mathbb{I}[x \ge 0],\ f_\nu(x) = kx^{k-1} e^{-x^k} \mathbb{I}[x \ge 0]$$

and

$$\mathrm{D}_\alpha(\eta \mid \nu) = \frac{1}{\alpha - 1} \log \int_0^\infty (f_\eta(x))^\alpha (f_\nu(x))^{1-\alpha}\, dx.$$

Note that

$$\begin{aligned}
&\int_0^\infty (f_\eta(x))^\alpha (f_\nu(x))^{1-\alpha}\,dx \\
&= \frac{k}{2^\alpha}\int_0^\infty x^{k-1-k\alpha/2}\exp\left(-\alpha x^{k/2}+(\alpha-1)x^k\right)dx \\
&\overset{y=x^{k/2}}{=} \frac{1}{2^{\alpha-1}}\int_0^\infty y^{1-\alpha}\exp\left(-\alpha y+(\alpha-1)y^2\right)dy \\
&= \infty.
\end{aligned}$$

Therefore, for $\alpha > 1$, $\mathrm{D}_\alpha(\eta \mid \nu) = \infty$.

Similarly,

$$\begin{aligned}
&\int_0^\infty (f_\nu(x))^\alpha (f_\eta(x))^{1-\alpha}\,dx \\
&= 2^{\alpha-1}k\int_0^\infty x^{k\alpha/2+k/2-1}\exp\left(-\alpha x^k - x^{k/2}+\alpha x^{k/2}\right)dx \\
&\overset{y=x^{k/2}}{=} 2^\alpha\int_0^\infty y^\alpha\exp\left(-\alpha y^2+(\alpha-1)y\right)dy < \infty.
\end{aligned}$$

Therefore, for $\alpha > 1$, $\mathrm{D}_\alpha(\nu \mid \eta) < \infty$. However,

$$\int_0^\infty |x-x'|^2\eta(dx) = \Gamma\left(1+\frac{4}{k}\right) - 2x'\Gamma\left(1+\frac{2}{k}\right) + (x')^2.$$

Minimizing this over x' gives us that the minimum is achieved at $x' = \Gamma\left(1+\frac{2}{k}\right)$. But

$$\lim_{k\searrow 0}\left[\Gamma\left(1+\frac{4}{k}\right) - 2\left(\Gamma\left(1+\frac{2}{k}\right)\right)^2 + \left(\Gamma\left(1+\frac{2}{k}\right)\right)^2\right] = \infty$$

and so $C_p^{\mathrm{PI}}(\eta) \nearrow \infty$ as $k \searrow 0$. □

Second, we work out the details of Example 5.2.

Proof of Example 5.2. Letting f_η and f_ν be the corresponding densities, we have

$$\begin{aligned}
&\int_{-\infty}^\infty (f_\eta)^\alpha (f_\nu)^{1-\alpha}\,dx \\
&= \frac{1}{(2\pi)^{\alpha/2}}\left(\frac{\Gamma\left((h+1)/2\right)}{\sqrt{h\pi}\,\Gamma(h/2)}\right)^{1-\alpha}\int_{-\infty}^\infty e^{-\alpha x^2/2}\left(1+\frac{x^2}{h}\right)^{(h+1)(\alpha-1)/2}dx. \\
&\le \frac{1}{(2\pi)^{\alpha/2}}\left(\frac{\Gamma\left((h+1)/2\right)}{\sqrt{h\pi}\,\Gamma(h/2)}\right)^{1-\alpha}\int_{-\infty}^\infty e^{-x^2/2}dx < \infty,
\end{aligned}$$

because, for $h \ge 2$, $\left(1+\frac{x^2}{h}\right)^{(h+1)(\alpha-1)/2} \le e^{(\alpha-1)x^2/2}$. Therefore, $\mathrm{D}_\alpha(\eta \mid \nu) < \infty$. However, for $h > 2$,

$$\int_{-\infty}^\infty |x-x'|^2\nu(dx) = \frac{h}{h-2} + (x')^2 \ge \frac{h}{h-2} \xrightarrow{h\searrow 2} \infty$$

and $C_p^{\mathrm{PI}}(\nu) \nearrow \infty$ as $h \searrow 2$. □

D. Proofs

D.1. Proof of Theorem 3.1

We begin by considering the case $d = 1$, dropping the component indexes from our notation.

Theorem D.1. *Assume $d = 1$. If $\mathcal{W}_1(\nu, \eta) \le \varepsilon$, then $|m_\nu - m_\eta| \le \varepsilon$ and $|\operatorname{MAD}_\nu - \operatorname{MAD}_\eta| \le 2\varepsilon$. On the other hand, if $\mathcal{W}_2(\nu, \eta) \le \varepsilon$, then $\mathcal{W}_1(\nu, \eta) \le \varepsilon$,*

$$|\sigma_\nu - \sigma_\eta| \le \frac{1}{2}\left(2^{1/2} + 6^{1/2}\right)\varepsilon \le 2\varepsilon,$$

and

$$\begin{aligned} |\sigma_\nu^2 - \sigma_\eta^2| &\le 2^{3/2}\min(\sigma_\nu, \sigma_\eta)\varepsilon + (1 + 3 \times 2^{1/2})\varepsilon^2 \\ &\le 3\min(\sigma_\nu, \sigma_\eta)\varepsilon + 5.25\varepsilon^2. \end{aligned}$$

The proof of Theorem D.1 is deferred to the next section. To generalize to the case of $d > 1$, for a random variable $\vartheta \sim \eta$ on $\mathbb{R}^d$ with distribution η and any vector $v \in \mathbb{R}^d$, let $m_{\eta,v} = \mathbb{E}(v^\top \vartheta)$, $\sigma_{\eta,v}^2 = \mathbb{E}\{(v^\top\vartheta - m_{\eta,v})^2\}$, and $\operatorname{MAD}_{\eta,v} = \mathbb{E}(|v^\top\vartheta - m_{\eta,v}|)$.

Corollary D.2. *Let $v \in \mathbb{R}^d$ satisfy $\|v\|_2 \le 1$. If $\mathcal{W}_1(\nu, \eta) \le \varepsilon$ then $|m_{\nu,v} - m_{\eta,v}| \le \varepsilon$ and $|\operatorname{MAD}_{\nu,v} - \operatorname{MAD}_{\eta,v}| \le 2\varepsilon$. On the other hand, if $\mathcal{W}_2(\nu, \eta) \le \varepsilon$, then*

$$\begin{aligned} &|\sigma_{\nu,v} - \sigma_{\eta,v}| \le \frac{1}{2}\left(2^{1/2} + 6^{1/2}\right)\varepsilon, \\ &|\sigma_{\nu,v}^2 - \sigma_{\eta,v}^2| \le 2^{3/2}\min(\sigma_{\nu,v}, \sigma_{\eta,v})\varepsilon + (1 + 3 \times 2^{1/2})\varepsilon^2. \end{aligned}$$

Proof. Let $\vartheta \sim \nu$, let $\vartheta_v = v^\top\vartheta_{(j)}$ and let ν_v denote the distribution of ϑ_v. Define $\hat{\vartheta}$, $\hat{\vartheta}_v$, and η_v analogously in terms of η. By the Cauchy-Schwarz inequality and the assumption that $\|v\|_2 \le 1$,

$$\mathbb{E}(|\vartheta_v - \hat{\vartheta}_v|^p) = \mathbb{E}(|v^\top\vartheta - v^\top\hat{\vartheta}|^p) \le \mathbb{E}(\|\vartheta - \hat{\vartheta}\|_2^p).$$

Hence $\mathcal{W}_p(\nu_v, \eta_v) \le \mathcal{W}_p(\nu, \eta)$. The corollary now follows from Theorem D.1. □

Lemma D.3. *For probability measures ξ, ν, η, we have $\|m_\nu - m_\eta\|_2 = \sup_{\|v\|_2 \le 1}|m_{\nu,v} - m_{\eta,v}|$, $\|\Sigma_\xi\|_2 = \sup_{\|v\|_2\le 1}\sigma_{\xi,v}^2$, and $\|\Sigma_\nu - \Sigma_\eta\|_2 = \sup_{\|v\|_2\le1}|\sigma_{\nu,v}^2 - \sigma_{\eta,v}^2|$.*

Proof. The first result follows since $m_{\nu,v} - m_{\eta,v} = v^\top(m_\nu - m_\eta)$ and for any $w \in \mathbb{R}^d$, $\sup_{\|v\|_2\le1} v^\top w = \|w\|_2$. For the second result, since Σ_ξ is positive semi-definite,

$$\|\Sigma_\xi\|_2 = \sup_{\|v\|_2\le1} v^\top\Sigma_\xi v = \sup_{\|v\|_2\le1}\mathbb{E}\{v^\top(X - m_\xi)(X - m_\xi)^\top v\} = \sup_{\|v\|_2\le1}\sigma_{\xi,v}^2;$$

The third result follows by an analogous argument. □

By taking $v = e_i$, the ith canonical basis vector of $\mathbb{R}^d$, Corollary D.2 implies the bounds in Theorem 3.1 on $|\operatorname{MAD}_{\nu,i} - \operatorname{MAD}_{\eta,i}|$ and $|\sigma_{\nu,i} - \sigma_{\eta,i}|$. Corollary D.2 and Lemma D.3 yield the bounds in Theorem 3.1 on $\|m_\nu - m_\eta\|_2$ and $\|\Sigma_\nu - \Sigma_\eta\|_2$.

D.2. Proof of Lemma 4.2

Proof. First, note that the $\mathsf{ELBO}(\xi)$ provides a lower bound for $\log M$ since $\operatorname{KL}(\xi \mid \pi) \ge 0$:

$$\begin{aligned} \mathsf{ELBO}(\xi) &:= \int \log\left(\frac{\mathrm{d}\pi'}{\mathrm{d}\xi}\right)\mathrm{d}\xi \\ &= \log M - \operatorname{KL}(\xi \mid \pi) \le \log M. \end{aligned} \tag{2}$$

Second, Jensen's inequality implies that $\mathsf{CUBO}_\alpha(\xi)$ is an upper bound for $\log M$:

$$\begin{aligned}\mathsf{CUBO}_\alpha(\xi) &:= \log\left\{\int\left(\frac{\mathrm{d}\pi'}{\mathrm{d}\xi}\right)^\alpha \mathrm{d}\xi\right\}^{1/\alpha} \\ &\geq \log\left\{\int \frac{\mathrm{d}\pi'}{\mathrm{d}\xi}\mathrm{d}\xi\right\} = \log M.\end{aligned}$$

The α-divergence is monotone in α, i.e., $\alpha \leq \alpha'$ implies that $\mathrm{D}_\alpha(\pi \mid \hat{\pi}) \leq \mathrm{D}_{\alpha'}(\pi \mid \hat{\pi})$ (Cichocki and Amari, 2010). Thus, by the definition of $\mathsf{CUBO}_\alpha(\hat{\pi})$ and Eq. (2), we have

$$\begin{aligned}\mathrm{KL}(\pi \mid \hat{\pi}) &= \mathrm{D}_1(\pi \mid \hat{\pi}) \leq \mathrm{D}_\alpha(\pi \mid \hat{\pi}) \\ &= \frac{\alpha}{\alpha-1}\left(\mathsf{CUBO}_\alpha(\hat{\pi}) - \log M\right) \\ &\leq \frac{\alpha}{\alpha-1}\left(\mathsf{CUBO}_\alpha(\hat{\pi}) - \mathsf{ELBO}(\eta)\right).\end{aligned}$$

□

Proof of Theorem D.1

Throughout we will always assume that $\vartheta \sim \nu$ and $\hat{\vartheta} \sim \eta$ are distributed according to the optimal coupling for the p-Wasserstein distance under consideration. We will also assume without loss of generality that $m_\nu = 0$ since if not we could consider the random variables $\vartheta' = \vartheta - m_\nu$ and $\hat{\vartheta}' = \hat{\vartheta} - m_\nu$ instead.

The 1-Wasserstein distance can be written as (Villani, 2009, Rmk. 6.5)

$$\mathcal{W}_1(\nu, \eta) = \sup_{\phi \,:\, \|\phi\|_L \leq 1} |\nu(\phi) - \eta(\phi)|. \tag{3}$$

By Jensen's inequality,

$$\mathcal{W}_q(\nu, \eta) \leq \mathcal{W}_p(\nu, \eta) \qquad (1 \leq q \leq p < \infty). \tag{4}$$

Eqs. (3) and (4) together imply that for any $p \geq 1$, if $\mathcal{W}_p(\nu, \eta) \leq \varepsilon$, then for any L-Lipschitz function ϕ, $|\nu(\phi) - \eta(\phi)| \leq L\varepsilon$.

Assume $\mathcal{W}_1(\nu, \eta) \leq \varepsilon$. By Eq. (3), for any Lipschitz function ϕ,

$$|\mathbb{E}(\phi(\vartheta) - \phi(\hat{\vartheta}))| \leq \varepsilon \|\phi\|_L.$$

Hence, taking $\phi(t) = t$, we have that $|m_\nu - m_\eta| = |m_\eta| \leq \varepsilon$. For the mean absolute deviation, using the fact that $\phi(t) = |t|$ is 1-Lipschitz, we have

$$|\,\mathrm{MAD}_\nu - \mathrm{MAD}_\eta\,| = |\mathbb{E}(|\vartheta| - |\hat{\vartheta} - m_\eta|)| \leq |\mathbb{E}(|\vartheta| - |\hat{\vartheta}|)| + |m_\eta| \leq 2\varepsilon.$$

Assume $\mathcal{W}_2(\nu, \eta) \leq \varepsilon$. By Jensen's inequality $\mathcal{W}_1(\nu, \eta) \leq \varepsilon$ as well. Let $\varsigma_\nu^2 = \mathbb{E}(\vartheta^2) = \sigma_\nu^2$ and $\varsigma_\eta^2 = \mathbb{E}(\hat{\vartheta}^2)$. It follows from the Cauchy-Schwarz inequality that

$$\begin{aligned}|\varsigma_\nu^2 - \varsigma_\eta^2| &= \left|\mathbb{E}\big(\vartheta^2 - \hat{\vartheta}^2\big)\right| = \left|\mathbb{E}\left\{\left(\vartheta - \hat{\vartheta}\right)\left(\vartheta + \hat{\vartheta}\right)\right\}\right| \\ &\leq \mathbb{E}\left\{\left(\vartheta - \hat{\vartheta}\right)^2\right\}^{1/2}\mathbb{E}\left\{\left(\vartheta + \hat{\vartheta}\right)^2\right\}^{1/2} \leq 2^{1/2}\varepsilon\mathbb{E}\big(\vartheta^2 + \hat{\vartheta}^2\big)^{1/2} \\ &\leq 2^{1/2}\varepsilon(\varsigma_\nu + \varsigma_\eta).\end{aligned}$$

Since $|\varsigma_\nu^2 - \varsigma_\eta^2| = |\varsigma_\nu - \varsigma_\eta|(\varsigma_\nu + \varsigma_\eta)$, it follows that

$$|\varsigma_\nu - \varsigma_\eta| \leq 2^{1/2}\varepsilon. \tag{5}$$

Using Eq. (5), we also have

$$|\sigma_\nu^2 - \sigma_\eta^2| = |\varsigma_\nu^2 - \varsigma_\eta^2 + m_\eta^2| \le |\varsigma_\nu^2 - \varsigma_\eta^2| + |m_\eta^2| \le 2^{1/2}\varepsilon(\varsigma_\nu + \varsigma_\eta) + \varepsilon^2 \tag{6}$$

$$|\sigma_\nu - \sigma_\eta| \le 2^{1/2}\varepsilon + \frac{\varepsilon^2}{\varsigma_\nu + \varsigma_\eta}. \tag{7}$$

If $\max(\sigma_\nu, \sigma_\eta) \le \frac{1}{2}\left(2^{1/2} + 6^{1/2}\right)\varepsilon$, then clearly $|\sigma_\nu - \sigma_\eta| \le \frac{1}{2}\left(2^{1/2} + 6^{1/2}\right)\varepsilon$. Otherwise $\frac{\varepsilon^2}{\varsigma_\nu + \varsigma_\eta} \le \frac{2\varepsilon}{2^{1/2}+6^{1/2}}$ and so, using Eq. (7), we have

$$|\sigma_\nu - \sigma_\eta| \le 2^{1/2}\varepsilon + \frac{2\varepsilon}{2^{1/2}+6^{1/2}} = \frac{1}{2}\left(2^{1/2} + 6^{1/2}\right)\varepsilon.$$

Hence we conclude unconditionally that $|\sigma_\nu - \sigma_\eta| \le \frac{1}{2}\left(2^{1/2} + 6^{1/2}\right)\varepsilon$. Starting with Eq. (6) and using Eq. (5), we have

$$\begin{aligned}|\sigma_\nu^2 - \sigma_\eta^2| &\le 2^{1/2}\varepsilon(\varsigma_\nu + \varsigma_\eta) + \varepsilon^2 = 2^{1/2}\varepsilon\left\{\sigma_\nu + (\sigma_\eta^2 + m_\eta^2)^{1/2}\right\} + \varepsilon^2 \\ &\le 2^{1/2}\varepsilon(2\sigma_\nu + 3\varepsilon) + \varepsilon^2 = 2^{3/2}\,\sigma_\nu\varepsilon + (1 + 3\times 2^{1/2})\varepsilon^2.\end{aligned}$$

D.3. Proof of Theorem C.1

Theorem D.4. *Let φ be a nonnegative measurable function on E and let $\delta > 0$. Then we have*

$$\|\varphi(\eta - \nu)\|_{TV} \le \left(27\left(\frac{1+\delta}{\delta}\right)^2 + 18 + 5\left(\log\int_E e^{\sqrt{2\varphi}}\,\mathrm{d}\nu\right)^2 + 3\mathrm{D}_{1+\delta}(\eta \mid \nu)^2\right)\mathrm{KL}(\eta \mid \nu)^{1/2}.$$

Corollary C.1 follows from Theorem D.4 and the fact that

$$\mathcal{W}_p(\eta, \nu)^p \le 2^{p-1}\|m(\vartheta', \cdot)^p(\eta - \nu)\|_{TV},$$

proved, for instance, in Villani (2003, Proposition 7.10). Indeed, it suffices to use $\varphi = \frac{\varepsilon^2}{2}m(\vartheta', \cdot)^p$ in Theorem D.4 to obtain the assertion.

D.4. Proof of Theorem D.4

We first assume, without loss of generality, that η is absolutely continuous with respect to ν, with density f. We set $u := f - 1$ so that

$$\eta = (1+u)\nu$$

and note that $u \ge -1$ and $\int_E u\mathrm{d}\nu = 0$. We also define

$$h(v) := (1+v)\log(1+v) - v, \quad v \in [-1, +\infty)$$

so that

$$\mathrm{KL}(\eta \mid \nu) = \int_E h(u)\mathrm{d}\nu. \tag{8}$$

We note that $h \ge 0$. We split the total variation in the following way:

$$\int \varphi\, d|\eta - \nu| = \int \varphi|u|\,\mathrm{d}\nu = \int_{\{-1\le u\le 4\}} \varphi|u|\,\mathrm{d}\nu + \int_{\{u>4\}} \varphi u\,\mathrm{d}\nu. \tag{9}$$

First part of the proof. In the first part, the first term ($u \le 4$) in (9) is bounded. This part is an adaptation of the first part of the proof of Bolley and Villani (2005, Theorem 1).

By Cauchy-Schwarz,

$$\int_{\{u\le 4\}} \varphi|u|\mathrm{d}\nu \le \left(\int_{\{u\le 4\}} \varphi^2\,\mathrm{d}\nu\right)^{1/2}\left(\int_{\{u\le 4\}} u^2\,\mathrm{d}\nu\right)^{1/2}.$$

On the other hand, from the elementary inequality,

$$-1 \le v \le 4 \quad \Longrightarrow \quad v^2 \le 4h(v)$$

(a consequence of the fact that $h(v)/v$ is nondecreasing), we deduce

$$\int_{\{u \le 4\}} u^2 \,\mathrm{d}\nu \le 4 \int_{\{u \le 4\}} h(u) \,\mathrm{d}\nu.$$

Combining this with the nonnegativity of h and (8), we find that

$$\int_{\{u\le 4\}} \varphi |u| \,\mathrm{d}\nu \le 2 \left(\int_E \varphi^2 \,\mathrm{d}\nu \right)^{1/2} \left(\int_E h(u) \,\mathrm{d}\nu \right)^{1/2} = 2 \left(\int_E \varphi^2 \,\mathrm{d}\nu \right)^{1/2} \mathrm{KL}(\eta \mid \nu)^{1/2}. \tag{10}$$

Now, since the function $t \mapsto \exp(\sqrt{2}\, t^{1/4})$ in increasing and convex on $\left[\frac{9^2}{2^2}, +\infty\right)$, we can write

$$\begin{aligned}
&\exp\left[\sqrt{2}\left(\int_E \varphi^2 \,\mathrm{d}\nu\right)^{1/4}\right] \\
&\le \exp\left[\sqrt{2}\left(\int_E \left(\varphi + \frac{9}{2}\right)^2 \mathrm{d}\nu\right)^{1/4}\right] \\
&\le \int_E \exp\left[\sqrt{2}\left(\left(\varphi + \frac{9}{2}\right)^2\right)^{1/4}\right] \mathrm{d}\nu \\
&= \int_E e^{\sqrt{2\varphi + 9}} \mathrm{d}\nu \\
&\le \int_E e^{\sqrt{2\varphi} + 3} \mathrm{d}\nu.
\end{aligned}$$

In other words,

$$\sqrt{2}\left(\int_E \varphi^2 \,\mathrm{d}\nu\right)^{1/4} \le 3 + \log \int_E e^{\sqrt{2\varphi}} \,\mathrm{d}\nu$$

and so

$$2\left(\int_E \varphi^2 \,\mathrm{d}\nu\right)^{1/2} \le \left(3 + \log \int_E e^{\sqrt{2\varphi}} \,\mathrm{d}\nu\right)^2.$$

Plugging this into (10), we conclude that

$$\int_{\{u\le 4\}} \varphi |u| \,\mathrm{d}\nu \le \left(3 + \log \int_E e^{\sqrt{2\varphi}} \,\mathrm{d}\nu\right)^2 \mathrm{KL}(\eta \mid \nu)^{1/2}. \tag{11}$$

Second part of the proof. Instead of following the logic of the second part of the proof of Bolley and Villani (2005, Theorem 1), which fails to provide the result we are seeking, we can note the following:

$$\begin{aligned}
\int_{u>4} \varphi u \,\mathrm{d}\nu &\le \frac{1}{(\log(5) - 1)^{1/2}} \int_{u>4} \varphi (u+1) \left(\log(u+1) - 1\right)^{1/2} \,\mathrm{d}\nu \\
&\le 2\left(\int_{u>4} \varphi^2 (u+1) \,\mathrm{d}\nu\right)^{1/2} \left(\int_{u>4} \left[(u+1)\left(\log(u+1) - 1\right) + 1\right] \mathrm{d}\nu\right)^{1/2} \\
&= 2\left(\int_{u>4} \varphi^2 \,\mathrm{d}\eta\right)^{1/2} \left(\int_{u>4} h(u) \,\mathrm{d}\nu\right)^{1/2} \\
&\le 2\left(\int_E \varphi^2 \,\mathrm{d}\eta\right)^{1/2} \mathrm{KL}(\eta \mid \nu)^{1/2}.
\end{aligned} \tag{12}$$

Now, since the function $t \mapsto \exp\left(\frac{\sqrt{2}\,\delta}{1+\delta} t^{1/4}\right)$ in increasing and convex on $\left[\frac{81(1+\delta)^4}{4\delta^4}, +\infty\right)$, we can write

$$\begin{aligned}
&\exp\left[\frac{\sqrt{2}\,\delta}{1+\delta}\left(\int_E \varphi^2 \,\mathrm{d}\eta\right)^{1/4}\right] \\
&\le \exp\left[\frac{\sqrt{2}\,\delta}{1+\delta}\left(\int_E \left(\varphi + \frac{9(1+\delta)^2}{2\delta^2}\right)^2 \mathrm{d}\eta\right)^{1/4}\right] \\
&\le \int_E \exp\left[\frac{\sqrt{2}\,\delta}{1+\delta}\left(\left(\varphi + \frac{9(1+\delta)^2}{2\delta^2}\right)^2\right)^{1/4}\right]\mathrm{d}\eta \\
&= \int_E e^{\sqrt{2\delta^2\varphi/(1+\delta^2)+9}}\mathrm{d}\eta \\
&\le \int_E e^{\delta\sqrt{2\varphi}/(1+\delta)+3}\mathrm{d}\eta.
\end{aligned}$$

In other words,

$$\frac{\sqrt{2}\,\delta}{1+\delta}\left(\int_E \varphi^2 \,\mathrm{d}\eta\right)^{1/4} \le 3 + \log\int_E e^{\delta\sqrt{2\varphi}/(1+\delta)}\,\mathrm{d}\eta$$

and so

$$2\left(\int_E \varphi^2 \,\mathrm{d}\eta\right)^{1/2} \le \left(\frac{1+\delta}{\delta}\right)^2 \left(3 + \log\int_E e^{\delta\sqrt{2\varphi}/(1+\delta)}\,\mathrm{d}\eta\right)^2.$$

Moreover, using Hölder's inequality,

$$\int_E e^{\delta\sqrt{2\varphi}/(1+\delta)}\,\mathrm{d}\eta \le \left(\int_E e^{\sqrt{2\varphi}}\,\mathrm{d}\nu\right)^{\delta/(1+\delta)} \left(\int_E f^{1+\delta}\mathrm{d}\nu\right)^{1/(1+\delta)}$$

and so

$$\begin{aligned}
\int_{u>4}\varphi u\,\mathrm{d}\nu &\le \left(\frac{1+\delta}{\delta}\right)^2\left(3 + \frac{\delta}{1+\delta}\log\int_E e^{\sqrt{2\varphi}}\,\mathrm{d}\nu + \frac{1}{1+\delta}\log\int_E f^{1+\delta}\mathrm{d}\nu\right)^2 \mathrm{KL}(\eta \mid \nu)^{1/2} \\
&\le \left(27\left(\frac{1+\delta}{\delta}\right)^2 + 3\left(\log\int_E e^{\sqrt{2\varphi}}\,\mathrm{d}\nu\right)^2 + 3\left(D_{1+\delta}(\eta|\nu)\right)^2\right)\mathrm{KL}(\eta \mid \nu)^{1/2}
\end{aligned}$$

Combining this with (11), we obtain the required result.

D.5. Proof of Theorem C.2

We have the following more general result, which we prove in the next section:

Theorem D.5. *Let φ be a nonnegative measurable function on E and let $q, q' > 1$ be such that $\frac{1}{q}+\frac{1}{q'} = 1$. Then we have*

$$\|\varphi(\eta-\nu)\|_{TV} \le \left[2\left(\int_E \varphi^2 \,\mathrm{d}\nu\right)^{1/2} + 2\left(\frac{1}{q}\int_E \varphi^{2q}\mathrm{d}\nu + \frac{1}{q'}\exp\left((q'-1)D_{q'}(\eta|\nu)\right)\right)^{1/2}\right]\mathrm{KL}(\eta \mid \nu)^{1/2}.$$

As described in more detail in Appendix D.3, Theorem C.2 follows immediately from Theorem D.5 when we use $\varphi = \frac{1}{2}m(\vartheta', \cdot)$.

D.6. Proof of Theorem D.5

We again assume, without loss of generality, that η is absolutely continuous with respect to ν, with density f. We set $u := f - 1$ so that

$$\eta = (1+u)\nu$$

and note that $u \geq -1$ and $\int_E u \mathrm{d}\nu = 0$. We also define

$$h(v) := (1+v)\log(1+v) - v, \quad v \in [-1, +\infty)$$

so that

$$\mathrm{KL}(\eta \mid \nu) = \int_E h(u) \mathrm{d}\nu.$$

We note that $h \geq 0$. We split the total variation in the following way:

$$\int \varphi \, d|\eta - \nu| = \int \varphi |u| \, \mathrm{d}\nu = \int_{\{-1 \leq u \leq 4\}} \varphi |u| \, \mathrm{d}\nu + \int_{\{u > 4\}} \varphi u \, \mathrm{d}\nu. \tag{13}$$

Using (10), we have that

$$\int_{\{u \leq 4\}} \varphi |u| \, \mathrm{d}\nu \leq 2 \left(\int_E \varphi^2 \, \mathrm{d}\nu \right)^{1/2} \mathrm{KL}(\eta \mid \nu)^{1/2}. \tag{14}$$

Furthermore, using (12), we have

$$\int_{u > 4} \varphi u \, \mathrm{d}\nu \leq 2 \left(\int_E \varphi^2 \, \mathrm{d}\eta \right)^{1/2} \mathrm{KL}(\eta \mid \nu)^{1/2}. \tag{15}$$

Using Young's inequality, we obtain

$$\int_E \varphi^2 \, \mathrm{d}\eta = \int_E \varphi^2 f \, \mathrm{d}\nu \leq \frac{1}{q} \int_E \varphi^{2q} \mathrm{d}\nu + \frac{1}{q'} \int_E f^{q'} \mathrm{d}\nu = \frac{1}{q} \int_E \varphi^{2q} \mathrm{d}\nu + \frac{1}{q'} \exp\left((q'-1) D_{q'}(\eta|\nu)\right),$$

which, together with (13), (14) and (15) gives the assertion.

D.7. Proof of Lemma 5.1

First assume that η and ν have densities f_η and f_ν with respect to Lebesgue measure. Note that the densities of the pushforward measures $T\#\eta$ and $T\#\nu$ are given by

$$x \mapsto f_\eta \circ T^{-1}(x) \left|\det J_x T^{-1}(x)\right| \quad \text{and} \quad x \mapsto f_\nu \circ T^{-1}(x) \left|\det J_x T^{-1}(x)\right|,$$

respectively, where J_x denotes the Jacobian. Therefore, for any $\alpha > 0$,

$$\begin{aligned}
\int \left(\frac{\mathrm{d}\,(T\#\eta)}{\mathrm{d}\,(T\#\nu)} \right)^\alpha \mathrm{d}\,(T\#\nu) &= \int \left(\frac{f_\eta \circ T^{-1}(x) \left|\det J_x T^{-1}(x)\right|}{f_\nu \circ T^{-1}(x) \left|\det J_x T^{-1}(x)\right|} \right)^\alpha f_\nu \circ T^{-1}(x) \left|\det J_x T^{-1}(x)\right| \mathrm{d}x \\
&= \int \left(\frac{f_\eta \circ T^{-1}(x)}{f_\nu \circ T^{-1}(x)} \right)^\alpha f_\nu \circ T^{-1}(x) \left|\det J_x T^{-1}(x)\right| \mathrm{d}x \\
&\overset{y = T^{-1}(x)}{=} \int \left(\frac{f_\eta(y)}{f_\nu(y)} \right)^\alpha f_\nu(y) \mathrm{d}y \\
&= \int \left(\frac{\mathrm{d}\eta}{\mathrm{d}\nu} \right)^\alpha \mathrm{d}\nu.
\end{aligned}$$

and so, for $\alpha \neq 1$, $\mathrm{D}_\alpha(\eta \mid \nu) = \mathrm{D}_\alpha(T\#\eta \mid T\#\nu)$. Similarly,

$$\begin{aligned}
\int \log \left(\frac{\mathrm{d}\,(T\#\eta)}{\mathrm{d}\,(T\#\nu)} \right) \mathrm{d}\,(T\#\eta) &= \int \log \left(\frac{f_\eta \circ T^{-1}(x)}{f_\nu \circ T^{-1}(x)} \right) f_\eta \circ T^{-1}(x) \left|\det J_x T^{-1}(x)\right| \mathrm{d}x \\
&= \int \log \left(\frac{f_\eta(y)}{f_\nu(y)} \right) f_\eta(y) \mathrm{d}y \\
&= \int \log \left(\frac{\mathrm{d}\eta}{\mathrm{d}\nu} \right) \mathrm{d}\nu.
\end{aligned}$$

and so $\mathrm{D}_1(T\#\eta \mid T\#\nu) = \mathrm{KL}(T\#\xi \mid T\#\pi) = \mathrm{KL}(\xi \mid \pi) = \mathrm{D}_1(\eta \mid \nu)$.

More generally, without assuming that η and ν are absolutely continuous with respect to Lebesgue measure, we note that if $\eta \ll \nu$ then

$$\frac{\mathrm{d}\,(T\#\eta)}{\mathrm{d}\,(T\#\nu)} = \frac{\mathrm{d}\eta}{\mathrm{d}\nu} \circ T^{-1}. \tag{16}$$

Indeed, for any measurable set A, we have

$$\int_A \frac{\mathrm{d}\eta}{\mathrm{d}\nu} \circ T^{-1}\, \mathrm{d}\,(T\#\nu) = \int_{T^{-1}(A)} \frac{\mathrm{d}\eta}{\mathrm{d}\nu} d\nu = (T\#\eta)\,(A).$$

Using Eq. (16) and the fact that T is bijective, we have that

$$\int \left(\frac{\mathrm{d}\,(T\#\eta)}{\mathrm{d}\,(T\#\nu)}\right)^{\alpha} \mathrm{d}\,(T\#\nu) = \int \left(\frac{\mathrm{d}\eta}{\mathrm{d}\nu} \circ T^{-1}\right)^{\alpha} \mathrm{d}\,(T\#\nu) = \int \left(\frac{\mathrm{d}\eta}{\mathrm{d}\nu}\right)^{\alpha} \mathrm{d}\nu. \tag{17}$$

Similarly,

$$\int \left(\frac{\mathrm{d}\log\,(T\#\eta)}{\mathrm{d}\,(T\#\nu)}\right) \mathrm{d}\,(T\#\eta) = \int \log\left(\frac{\mathrm{d}\eta}{\mathrm{d}\nu} \circ T^{-1}\right) \mathrm{d}\,(T\#\eta) = \int \log\left(\frac{\mathrm{d}\eta}{\mathrm{d}\nu}\right) \mathrm{d}\eta. \tag{18}$$

Eq. (17) and Eq. (18) prove that $\mathrm{D}_\alpha(\eta \mid \nu) = \mathrm{D}_\alpha(T\#\eta \mid T\#\nu)$ for any $\alpha > 0$.

D.8. Proof of Proposition 5.3

As described in more detail in Appendix D.3, Proposition 5.3 follows immediately from the following result:s

Theorem D.6. *Let φ be a nonnegative measurable function on E and suppose that η and ν are probability measures and $\eta \ll \nu$. Then*

$$\|\varphi(\eta-\nu)\|_{TV} \le \left(\int \varphi^2\, d\nu\right)^{1/2} \left(\exp\left\{D_2(\eta|\nu)\right\} - 1\right)^{1/2}.$$

Proof. Let $f = \frac{\mathrm{d}\eta}{\mathrm{d}\nu}$. We set $u := f - 1$ so that

$$\eta = (1+u)\nu.$$

Note that the total variation can be expressed in the following way

$$\begin{aligned}
\int \varphi\, d|\eta-\nu| &= \int \varphi |u|\, \mathrm{d}\nu \\
&\le \left(\int \varphi^2\, \mathrm{d}\nu\right)^{1/2} \left(\int u^2\, \mathrm{d}\nu\right)^{1/2} \\
&\le \left(\int \varphi^2\, \mathrm{d}\nu\right)^{1/2} \left(\int (f^2 - 2f + 1)\, \mathrm{d}\nu\right)^{1/2} \\
&= \left(\int \varphi^2\, \mathrm{d}\nu\right)^{1/2} \left(\int f^2\, \mathrm{d}\nu - 1\right)^{1/2} \\
&= \left(\int \varphi^2\, \mathrm{d}\nu\right)^{1/2} \left(\exp\left\{D_2(\eta|\nu)\right\} - 1\right)^{1/2}.
\end{aligned}$$

□